

CRESST REPORT 763

Julia Phelan
Kilchan Choi
Terry Vendlinski
Eva L. Baker
Joan L. Herman

THE EFFECTS OF POWERSOURCE[®]
INTERVENTION ON STUDENT
UNDERSTANDING OF BASIC
MATHEMATICAL PRINCIPLES

DECEMBER, 2009



National Center for Research on Evaluation, Standards, and Student Testing

Graduate School of Education & Information Studies
UCLA | University of California, Los Angeles

**The Effects of POWERSOURCE[®] Intervention on
Student Understanding of Basic Mathematical Principles**

CRESST Report 763

Julia Phelan, Kilchan Choi, Terry Vendlinski, Eva L. Baker, and Joan L. Herman
CRESST/University of California, Los Angeles

December, 2009

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

Copyright © 2009 The Regents of the University of California

The work reported herein was supported under the National Research and Development Centers, PR/Award Number R305A050004, as administered by the U.S. Department of Education, Institute of Education Sciences.

The findings and opinions expressed in this report do not necessarily reflect the positions or policies of the National Research and Development Centers of the U.S. Department of Education, Institute of Education Sciences.

To cite from this report, please use the following as your APA reference:

Phelan, J., Choi, K., Vendlinski, T., Baker, E. L., & Herman, J. L. (2009). *The effects of POWERSOURCE[®] intervention on student understanding of basic mathematical principles* (CRESST Report 763). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

THE EFFECTS OF POWERSOURCE[®] INTERVENTION ON STUDENT UNDERSTANDING OF BASIC MATHEMATICAL PRINCIPLES

Julia Phelan, Kilchan Choi, Terry Vendlinski, Eva L. Baker, & Joan L. Herman
CRESST/University of California, Los Angeles

Abstract

This report describes results from field-testing of POWERSOURCE[®] formative assessment alongside professional development and instructional resources. The researchers at the National Center for Research, on Evaluation, Standards, & Student Testing (CRESST) employed a randomized, controlled design to address the following question: Does the use of POWERSOURCE[®] strategies improve 6th-grade student performance on assessments of the key mathematical ideas relative to the performance of a comparison group? Sixth-grade teachers were recruited from 7 districts and 25 middle schools. A total of 49 POWERSOURCE[®] and 36 comparison group teachers and their students (2,338 POWERSOURCE[®], 1,753 comparison group students) were included in the study analyses. All students took a pretest of prerequisite knowledge and a transfer measure of tasks drawn from international tests at the end of the study year. Students in the POWERSOURCE[®] group used sets of formative assessment tasks. POWERSOURCE[®] teachers had exposure to professional development and instructional resources. Results indicated that students with higher pretest scores tended to benefit more from the treatment as compared to students with lower pretest scores. In addition, students in the POWERSOURCE[®] group significantly outperformed control group students on distributive property items and the effect was larger as pretest scores increased. Results, limitations and future directions are discussed.

Introduction

Formative assessment has two valuable features: First it can provide teachers with feedback on the effectiveness of their teaching strategies, which in turn allows teachers to adjust their instructional strategies or sequence to improve student learning. Second, formative assessment can also be useful for students, helping them to identify the areas to which they must devote time or effort, and whether they need to adjust their thinking in any way. There is evidence that learning gains in classrooms using formative assessment can be substantial—reviews suggest one-half to one full standard deviation increase (Black and Wiliam, 1998a). This finding holds true across multiple age groups, school subjects and countries and is particularly robust for low achieving students (Wiliam, Lee, Harrison, &

Black, 2004). In other words, when efforts are made to strengthen formative assessment, learning gains (typically measured by an increase in test scores) occur.

Although research clearly highlights the promise of formative assessment, this promise may be hard to fulfill. Researchers can agree that formative assessment is a concrete way for teachers to improve learning efficiency in their classrooms, however, implementing a formative assessment program in the classroom may prove challenging. The challenges arise from both the uneven quality of available formative assessments and the limitations of teachers' capacity. For example, data on the quality of interim or benchmark testing, marketed as formative assessments to districts and schools is scarce and assessments included with curriculum tends to be an afterthought rather than a core, quality element of current materials (Herman & Baker, 2006; Herman, Osmundson, Ayala, Schneider, & Timms, 2006; Wolf, Bixby, Glenn, & Gardner, 1991). Moreover, educators often have limited background and capacity to develop or engage in quality assessment practices (Heritage & Yeagley, 2005; Herman & Gribbons, 2001; Plake & Impara, 1997; Shepard, 2001; Stiggins, 2005). For formative assessment to be effective, it must be accompanied by clear criteria and high-quality feedback. And information about the assessments must be delivered to the right people at the right time. There must also be the opportunity for the student to receive timely feedback from the assessor and this communication must be clear enough that the student understands what they can do to improve.

Both anecdotal and research evidence from districts across the U.S. suggests that many teachers are unable to use the information from benchmark tests or their own assessments because they lack the knowledge, materials, or curricular time to do so. As a result, there is a great deal of rhetoric surrounding formative assessment and "doing something" with the results, but in reality teachers don't always have the wherewithal to do anything except repeat what they have already done. As Stiggins (2004) notes, "teacher must possess and be ready to apply knowledge of sound classroom assessment practices...if teachers assess accurately and use the results effectively, then students will prosper," (p. 26). Thus teachers' pedagogical content knowledge and subject matter knowledge poses a challenge when it comes to executing a formative assessment plan.

Learning to use assessment in a more formative way often requires significant changes on the part of districts, teachers, and students. For districts, this change may mean insuring that teachers have both the time and resources to act on the assessment information they receive. For teachers and students, it involves learning to use assessment information diagnostically to determine the course of instruction and learning, and to deal with learning difficulties that are revealed by the assessments.

Responding to the Formative Assessment Challenge

The POWERSOURCE[®] strategy was developed in response to the need for good quality formative assessments in the classroom. In particular we wanted to address the joint challenges of assuring high quality formative assessments and of enabling teachers to use formative assessments more effectively and efficiently. Based on research on learning and targeting fundamental principles (or big ideas) of middle-school mathematics, POWERSOURCE[®] includes both a system of learning-based assessments and an infrastructure to support teachers' use of those assessments to improve student learning. The specific purpose of this strategy is to provide assessment information and resources to middle-school teachers, with the aim of improving both teachers' and students' understanding of the key ideas that are the prerequisites to mastering algebra. The intervention is predicated on the assumption that effective formative assessment must include not just validated assessments (Phelan, Kang, Niemi, Vendlinski, & Choi, 2009) but also instructional strategies and resources linked to the assessments, as well as professional development to make sure that teachers know how to use information from the assessments.

In our initial studies we sought to test the strategy and determine whether teachers could use, short, targeted assessments (along with an infrastructure to support their use) and if using these assessments would have an impact on student learning. Results were positive and indicated that teachers could use short formative assessments effectively and student gains were found (Phelan, Choi, Vendlinski, Baker, & Herman, 2009). Specifically, when compared to a comparison group, students in the POWERSOURCE[®] group performed better on both extended response and short-answer questions presented within the formative assessment framework. These findings demonstrated both the feasibility and value of including performance task-types in a brief assessment context.

The POWERSOURCE[®] Intervention

The POWERSOURCE[®] intervention focuses on middle-school mathematics, starting in 6th grade, and on helping assure that students possess key understandings they need for success in Algebra 1. The focus on algebra is motivated by ample research showing the frequency and price of failure for subsequent academic performance, including high school graduation, college entry, and preparation (e.g., Brown & Niemi, 2007). For example, data from the California State Algebra I exam over the past 5 years reveals that on average, 76% of students are below proficiency (California Department of Education, 2008). Although the current focus is on algebra in middle school, in the future, we expect to apply the POWERSOURCE[®] strategy across curriculum and across grade levels.

The POWERSOURCE[®] intervention targets big ideas and related skills in four domains underlying success in Algebra 1: (a) rational number equivalence (RNE), (b) properties of arithmetic (PA; the distributive property), (c) principles for solving linear equations (SE); and (d) application of core principles in these domains to other critical areas of mathematics, such as geometry and probability (RA). These domains were chosen because of their importance to later mastery of algebra and their significant place in state mathematics standards across Grades 6–8.

Item development procedures drew on assessment models we have validated in an extensive series of studies over many years (e.g., Baker, Freeman, & Clayton, 1991; Niemi, 1996; Baker, 1997; Niemi, Sylvester, & Baker, 2007), capitalizing on the idea that assessment developers work with specific architectures that are designed to elicit and evaluate student responses consistent with intended cognitive demands and to be applied in specific subject matter content. Using these architectures, assessment development teams composed of mathematics educators and experienced item writers from CRESST developed an initial pool of items; these were reviewed for content relevance and potential bias by mathematics educators and CRESST staff; then items passing this review were piloted with 6th-grade students. During the 2005–2006 school year, a large pool of formative assessment items that represented the content domain and cognitive demands of interest and forms were designed and tested. These items included basic computational tasks, partially worked examples, word problems, graphics problems, and explanation tasks. Results of the pilot studies were used to refine the item set and select items for inclusion in the larger field test describe here (for more details on item development and validation see Phelan, Kang, et al., 2009).

Within each of the selected content areas we designed a series of short POWERSOURCE[®] assessments called *Checks for Understanding* to help 6th-grade teachers assess their students' understanding of basic mathematical principles and to connect their instruction and provide feedback to support deeper understanding. For example, Figure 1 shows some items from one of the Checks for Understanding of RNE. These items were intended to elicit students' understanding of how to find equivalent rational numbers, how one can use the multiplicative identity property to help find equivalent rational numbers, what rational number equivalence is, and procedures for finding equivalent numbers by using the multiplicative identity.

5 Explain why the following three fractions are equivalent.

$\frac{2}{2}$ $\frac{100}{100}$ $\frac{1171}{1171}$

1 $75 \cdot 1 =$


2 Name and explain the property that you used to find the answer to question 1.

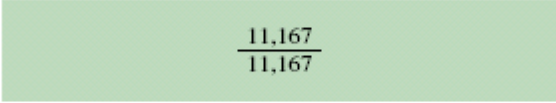
Figure 1. Assessment items from one of the Checks for Understanding focused on rational number equivalence.

During our pilot testing periods and through initial meetings with teachers, it became clear to us that teachers would benefit from some additional instructional supports to accompany the Checks for Understanding. We created a set of materials for teachers to help them teach the content, and also score and understand the results of the assessments. We designed these materials for teachers to use as support when teaching each of the domains addressed in the study. Working with expert teachers from one of our participating districts, we developed four teacher handbooks—each one was closely aligned with one of the four Checks for Understanding domains (see Figure 2 for an example page from an instructional handbook).

Lead discussion about the fact that any non-zero number over itself equals 1.


Then ask students for other examples of fractions equal to 1. Draw a large 1 around the fractions.

draw 

write 

say What about 11,167 over 11,167?
Does this fraction equal 1?

Then give students some fractions with only a numerator or denominator and ask how to make these fractions equal to 1. For example:

write 

say So now you know what a rational number is. You know that fractions are rational numbers and you know how to find fractions that are equal to 1.

Figure 2. Example page from instructional handbook on rational number equivalence.

These handbooks were designed as guides for each of two lessons. The handbooks contained a lesson “script” for the first day of instruction and a student worksheet made up the second day of instruction. The lesson script could be used exactly as it appeared in the book, or teachers could use the concepts and/or examples therein to augment their own teaching. The important thing was for teachers to make sure they covered all the concepts in the script and use the examples provided within. The student worksheet included as the second day’s instruction, could be used either as a whole class, teacher-directed activity, as a guided-instruction worksheet, or could be the basis of a group/pair working session. How teachers decided to use the instruction for the second day depended on how well students performed on the first assessment. Teachers were given simple forms on which to tally how many students missed each problem on the first assessment. This information could then be used, in an expeditious way, to inform how to teach the second day’s content.

Professional development was another component of the POWERSOURCE[®] intervention. The professional development consisted of an initial meeting during which teachers were given an overview of the study objectives and the theoretical underpinnings of the project. It was important that they understood the content of the big ideas as well as the formative assessment process, which was new to some of them. Teachers were also given advice on how to look at and use student data to gather information on student understanding and to change instruction. Three follow-up meetings (after each of the first three instructional modules) were also a part of the POWERSOURCE[®] professional development. At these follow-up meetings, teachers had the opportunity to look at student assessment data from within their district. Data were analyzed for the district as a whole and comparisons drawn amongst the percentages of correct vs. incorrect responses and so on. Looking at response patterns is something teachers reportedly did not get the opportunity to do very often. Typically, results from interim assessments reported on percentages right or wrong, but gave no additional information on what the most common incorrect responses were. These efforts are described in more detail elsewhere (Howard, Vendlinski, Hemberg, Niemi, & Phelan 2009).

Thus, a POWERSOURCE[®] module around a given domain (e.g., rational number equivalence) included a set of Checks for Understanding, targeted instructional resources, and professional development.

The POWERSOURCE[®] Sequence

The POWERSOURCE[®] materials were designed to complement existing curricula, but time for them must be found within tight district curriculum frameworks and timelines. It was therefore important for us to develop and pilot test formative assessments that would integrate well and easily with existing initiatives and not add an unreasonable burden to the heavy testing requirements already imposed on teachers (e.g., weeks of state and district testing), and not replace large chunks of extant curricula. Because the four domains that POWERSOURCE[®] addresses are treated, at least to some extent, in every 6th-grade mathematics program, POWERSOURCE[®] assessments can be easily incorporated at appropriate points into any ongoing curriculum. Checks for Understanding were revised based on data from the 2006–2007 study. Moreover, three Checks for Understanding were developed for each of the four 6th-grade POWERSOURCE[®] domains—instead of the two used the previous year. The initial Check for Understanding consisted of between 8–10 items and was given prior to instruction in the relevant content. This Check for Understanding acted as a baseline assessment for each student and (a) gave teachers information about their students initial knowledge and (b) allowed us to compare students' performance before and

after instruction, thus providing information on the instructional sensitivity of the Checks for Understanding. Each subsequent Check for Understanding (of which there are two) consisted of 4–5 items (2 symbolic representation/computation items and 2–3 open-ended problem solving and/or explanation tasks). Based on our research and development over the last 2 years, teachers’ procedure for using the Checks for Understanding was as follows:

1. Administer an initial Check for Understanding a big idea and its applications (15–20 minutes), analyze results.
2. Present instructional activities (if necessary) addressing deficiencies in conceptual understanding identified in step 1 (one class period).
3. Administer a second Check for Understanding focusing on conceptual understanding (15 minutes), and follow up instruction if necessary.
4. Present instruction on applications of the big idea to problem solving and symbolic representation and computation tasks (if necessary) (15 minutes).
5. Administer a third Check for Understanding focusing on conceptual understanding (15 minutes), and follow up instruction if necessary.

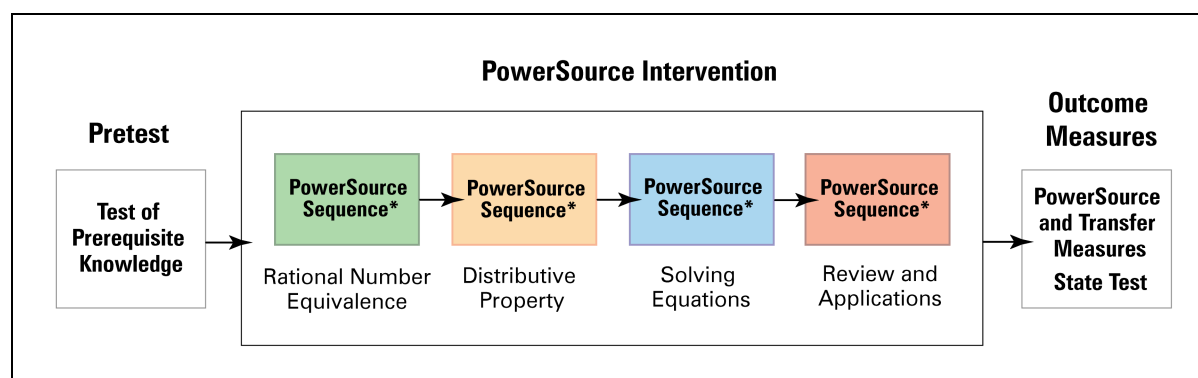


Figure 3. Overview of the Sequence of POWERSOURCE® Assessments

Instructional resources for teachers included guidance on understanding student responses to the POWERSOURCE® assessments; teaching activities and sample scripts to help students grasp the key concepts and use them to solve complex problems; explanations, models, and demonstrations of the meaning of the core concepts; worked examples for teaching problem solving schemas; and professional development focused on using the assessments and instructional resources effectively and efficiently.

The core undertaking of our work described here was conducting an extended, random assignment implementation study of our 6th-grade POWERSOURCE® program. Teachers were randomly assigned to either POWERSOURCE® or control conditions with the ultimate goal of determining program impact on both students and teacher learning outcomes. The

treatment group students in our POWERSOURCE[®] study received instruction and formative assessments (Checks for Understanding) on the four POWERSOURCE[®] domains. Also included in the study were a control group of students who received their regular instruction.

We hypothesize that students in the POWERSOURCE[®] group would possess a better understanding of the basic mathematical principles contained within each domain. We also hypothesize that students will be able to apply concepts they have learned, solve complex problems and transfer the principles covered by the POWERSOURCE[®] domains. For example, having received instruction and formative assessment on rational number equivalence, students should understand the multiplicative identity principle and be able to use it to: (a) demonstrate that a set of rational numbers are equivalent, (b) find equivalent fractions, (c) find missing numbers in proportions, and (d) solve proportional reasoning problems.

Study Sample

We spent considerable time and effort to secure additional school district participation in the project for the 2007–2008 school year, meeting with representatives of close to 20 districts. Ultimately, 7 school districts participated in the random assignment implementation study. As compared to larger districts, smaller districts (such as the ones included in our study) tended to have fewer algebra initiatives already in progress. Often they were also more open to participating in an experimental study so that not all teachers in the district needed to use the same program and materials. The latter issue was a stumbling point in discussions with several large districts; administrators and other personnel were interested in being involved, but only to the extent that all teachers could participate—thus precluding a randomized-controlled study at these sites. All of the districts reported in this study followed our randomized design. Our original sample was 91 sixth-grade teachers from 7 school districts. Table 1 shows the initial distribution of teachers in each district.

Table 1

Initial Sample Distribution by School District (2007–2008 school year)

District	<i>N</i> of students	<i>N</i> of teachers	<i>N</i> of schools	Design
AZ-1	93	2	1	B-S
CA-1	872	18	3	W-S
CA-2	770	11	2	W-S
CA-3	195	5	3	B-S
CA-4	279	11	5	B-S
CA-5	974	20	5	W-S
CA-6	1,232	24	8	B-S

Note. AZ = Arizona, CA = California, B-S = between school, W-S = within school.

Although the inclusion of a larger number of districts than originally planned was driven by practical considerations, methodologically it was also a definite strength, providing us the opportunity to see how well the program worked in a variety of settings. The design also allowed us to investigate factors that might magnify or dampen POWERSOURCE[®] effects, adding the qualities of a rich, mixed methods multi-site case study. With one large district, we might get a very precise estimate of the treatment, but it might pertain to only one set of district conditions.

Due to administrative reasons, many pretest scores in AZ-1 district were not valid and therefore were excluded from the analyzed sample as missing values. Accordingly, the number in the sample in this district was substantially smaller compared to the numbers in the other school districts. Also, we had to remove two of the schools from our initial between-school (B-S) design sample owing to issues of noncompliance with study procedures. Our final sample was 85 teachers from 27 schools in the 7 school districts participated in the study (see Table 2).

Table 2

Sample Distribution by School District (2007–2008 school year)

District	<i>N</i> of students	<i>N</i> of teachers	<i>N</i> of schools	Design
AZ-1	93	2	1	B-S
CA-1	872	18	3	W-S
CA-2	770	11	2	W-S
CA-3	195	5	3	B-S
CA-4	279	11	5	B-S
CA-5	974	20	5	W-S
CA-6	908	18	6	B-S

Note. AZ = Arizona, CA = California, B-S = between-school, W-S = within-school.

Student characteristics for participating schools are presented in Table 3. In all four districts, the percentage of students who were below proficiency in math the previous year was 55%. The percentage of English learner students ranged from 10 to 26%.

Table 3

Student Characteristics for Participating Schools

Student characteristics:	AZ District 1	CA District 1	CA District 2	CA District 3	CA District 4	CA District 5	CA District 6
Asian	0%	6%	4%	1%	6%	12%	3%
Black	10%	3%	7%	1%	6%	13%	3%
Hispanic	33%	36%	41%	24%	70%	32%	76%
White or other	57%	55%	52%	74%	16%	13%	18%
EL	12%	12%	14%	10%	22%	26%	20%
Below proficient in math, 2007	24%	50%	50%	49%	55%	53%	49%

Note. AZ = Arizona, CA = California, EL = English learner.

Study Design

Prior field-testing concentrated on a smaller number of districts (4) using within-school (W-S) designs (Phelan, Kang, et al., 2009). Based on district needs and configuration we incorporated both within- and between-school random assignment models. Ultimately, three of the districts used a W-S design, where random assignment was accomplished within each school (i.e., a given school had both POWERSOURCE[®] and control teachers). Four districts

used a B-S design, where schools within a district were randomly assigned to either the POWERSOURCE[®] or the control condition (see Table 4; also see Table 1 for design implemented in each district).

Table 4
Sample Distribution ('07–'08 school year)

Design	Control/Treatment	<i>N</i> of students	<i>N</i> of teachers	<i>N</i> of schools
Between	Control	633	13	8
	Treatment	842	23	7
	Subtotal	1475	36	15
Within	Control	1120	23	5
	Treatment	1496	26	5
	Subtotal	2616	49	10

Although the content focus of the four POWERSOURCE[®] modules remained the same (RNE, PA, SE, and RA), we changed the structure of each unit somewhat based on teacher feedback and our implementation experiences during the field test year. In the current study, POWERSOURCE[®] teachers were provided with three Checks for Understanding for each unit—one prior to the first day's set of instructional materials, one in between the first and second day of instruction, and one after the second day of instruction. Students in the control group did not complete any of the Checks for Understanding. Thus, the control students and teachers had no exposure to any of the POWERSOURCE[®] materials or concepts during the school year. All students (POWERSOURCE[®] and control) completed a test of prerequisite knowledge at the beginning of the school year and transfer measures of math knowledge at the end of the school year (described below). Based on district response and feedback, control teachers were offered the option of an alternative (i.e., non-POWERSOURCE[®]) professional development CRESST program, as opposed to standard district professional development. The majority of participating districts selected this alternative, non-POWERSOURCE[®] professional development.

Measures

In addition to POWERSOURCE[®] revisions based on feedback from the prior year's pilot testing, we refined the project outcome measures, introducing a pretest and an end of the year transfer measure. To investigate the quality of the test items, we used the one parameter logistic model (1PLM) for dichotomous items and partial credit model (PCM;

Masters, 1982) for polytomous items. This choice is in line with findings in Phelan et al. (2009) which showed the appropriateness in using unidimensional Rasch models for POWERSOURCE[®] test items. Additionally, with the data sets for the pretest and transfer measure having items representing all domains, factor analysis was conducted to estimate the amount of variance explained by the main construct.

POWERSOURCE[®] Pretest. The test of prerequisite knowledge served as a baseline measure for later analyses. The pretest consisted of 28 items modeled after items on the California State Test released items for 5th grade, and items used and validated on other CRESST projects. These items addressed concepts covered within the scope of the POWERSOURCE[®] research.

Item Analyses. Among the 28 items on the 2007–2008 POWERSOURCE[®] pre-test, the item PRE04 was the easiest item ($b = -2.600$, p -value = 0.99), and PRE23 was the most difficult item ($b = 3.139$, p -value = 0.13; see Figure 4 for these items).

4. Solve: $4 + \square = 10$

a) 8
b) 10
c) 4
d) 6

Item PRE04

23. $(4 \div 6) \div 2$ has the same value as:

a) $4 \cdot 6 \cdot 2$
b) $4 \div 6 \cdot \frac{1}{2}$
c) $4 \div (6 \div 2)$
d) $4 \div 6 \cdot 2$

Item PRE23

Figure 4. Easiest and most difficult items on the 6th-grade Pretest.

The polyserial correlation coefficients between item and test scores were larger than 0.3 except for two items (PRE23 and PRE24). These two appeared to have poor discrimination, and were deemed poor quality items. The test and item reliability based on item response theory were calculated. The IRT test reliability was calculated as Dimitrov (2003) suggested and it was .917 (Cronbach's alpha = .80). And, the most difficult item (PRE23) had the smallest item reliability, which means it has less contribution to test reliability than the other items in the pretest.

The item information curves in Figure 5 show that PRE04 and PRE08 mainly give information for the examinees with low ability. Difficult items PRE27 and PRE23 provide relatively large amounts of information for the examinees with high ability.

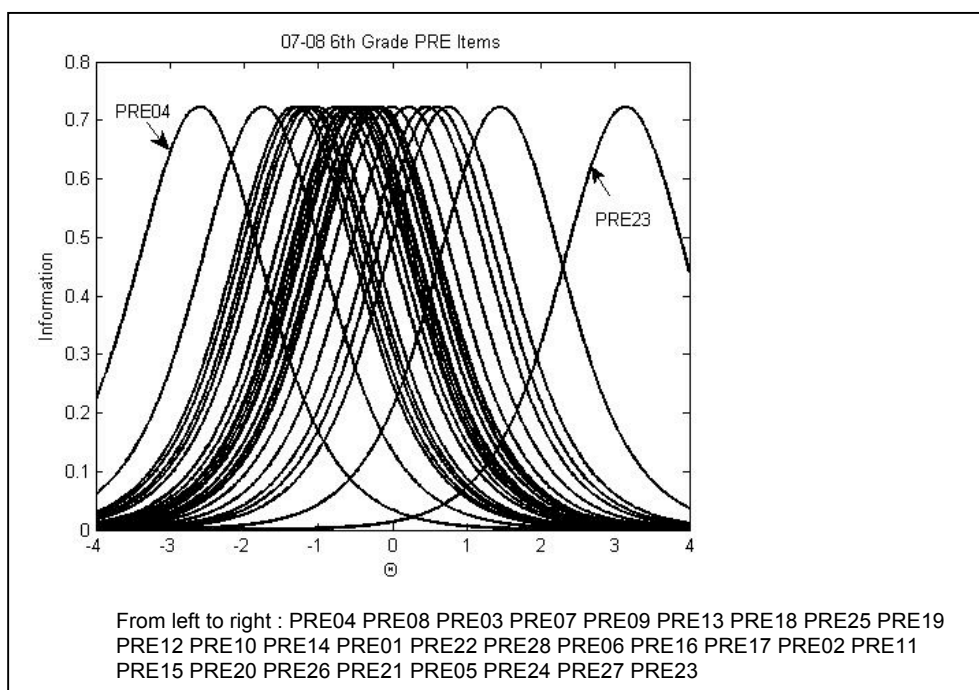


Figure 5. The item characteristic curves of POWERSOURCE® pretest items.

POWERSOURCE® Transfer Measure. Although we saw strong effect sizes in the prior year's field test based on POWERSOURCE® measures (see Phelan et al., 2009), we recognized the need for demonstrating intervention effects on an independent, transfer measure. Consequently, we developed a transfer measure using items from several sources including the Third International Mathematics and Science Study (TIMSS), National Assessment of Educational Progress (NAEP), the Qualifications and Curriculum Authority (QCA) Key Stage 3 exam, Programme for International Student Assessment (PISA) and benchmark tests used in one of our pilot districts. An initial set of 44 items were selected

from the various sources. Items were selected based on their relevance to the POWERSOURCE[®] domains and their appropriateness for a transfer task (i.e., related to POWERSOURCE[®] content, but not exact replicas of item types used in the *Checks for Understanding*). A final set of items (29) were selected from the initial 44 items. Of these items 19 were multiple choice, 9 short answer and 1 explanation task. Items were selected based on their representation in the California state standards and relevance to POWERSOURCE[®] items. Some of the initially developed items were deemed more appropriate for 7th grade and will be used for the 7th-grade transfer measure. The transfer measure was given to all participating students (POWERSOURCE[®] and control) at the culmination of the study year ($N = 5,358$).

Item Analyses. Item analyses were carried out on 29 transfer measure items (there were 31 possible responses as 1 item had 3 parts). Among the 31 items (30 dichotomous and 1 polytomous), we determined degree of difficulty from easiest ($b = -1.629$, p -value = 0.911) to most difficult ($b = 1.296$, p -value = 0.141). The polyserial correlation coefficients between item and test scores were larger than 0.3 except for one item that had poor discrimination. The test and item reliability based on item response theory were also calculated. The test reliability was .93 (Cronbach's alpha = .86). The polytomous item (POST27) had the largest item reliability, which means it has more contribution to test reliability than the other items in transfer measure. The item information curves also indicated that the explanation task provided the largest amount of information (see Figure 6).

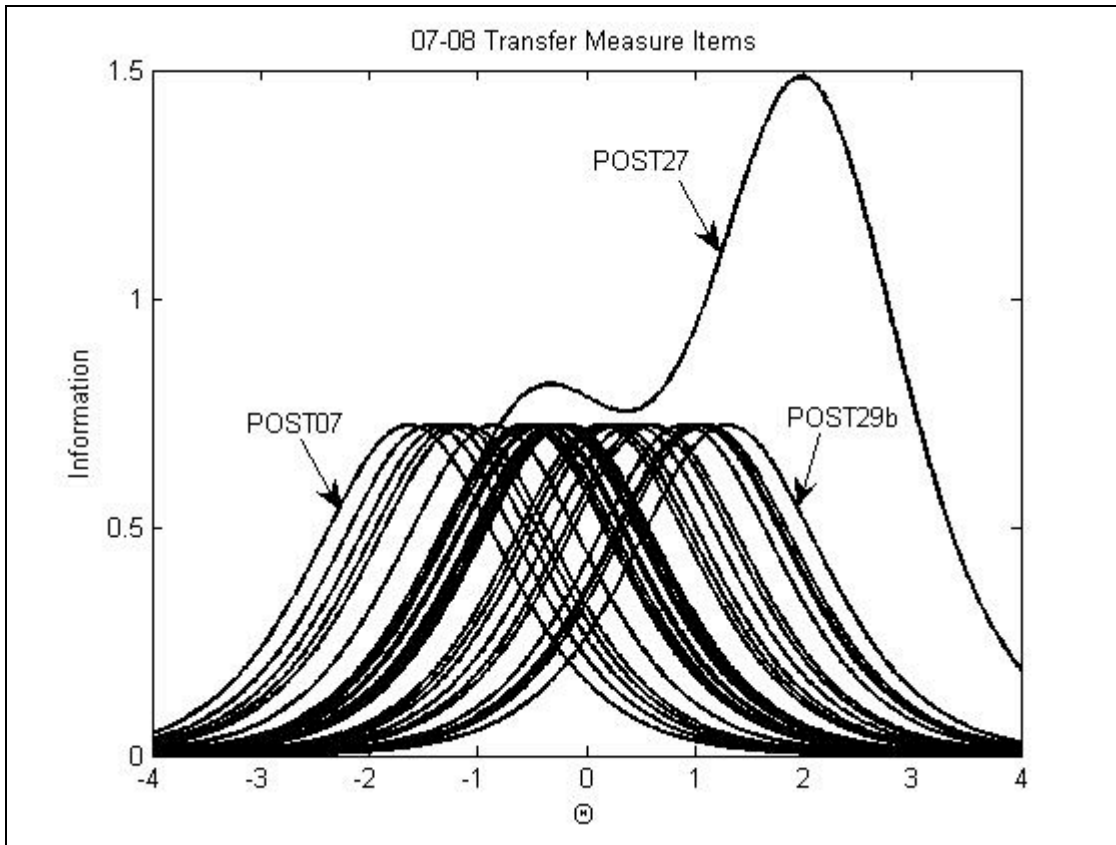


Figure 6. The item characteristic curves of POWERSOURCE© Transfer Measure items.

As shown in the Figure 6, the easiest item (POST07) mainly gives information for the examinees with low ability and the most difficult item (POST29b) is providing more information for the examinees with high ability. See Appendix C for the complete set of transfer measure items used.

Reliability

Table 5 shows the number of items, the actual number of examinees, and reliability for both the pretest and the transfer measure. The reliability was computed with coefficient alpha as shown in Table 5. The reliability coefficient for the pretest was 0.8 and for the transfer measure 0.86.

Table 5

Sample Size and Reliability of the 2007–2008 POWERSOURCE© Assessments

POWERSOURCE© assessments	# Items (= # dichotomous items + # polytomous items)	Sample size	Reliability (Cronbach's alpha)
Pre-test	28 (= 28 + 0)	5,838	.80
Transfer measure (Post-test)	31 (= 30 + 1)	5,358	.86

Descriptive Statistics

We first present descriptive statistics for the pretest score and transfer measure. Students who are missing either a pretest or a transfer measure score were treated as missing data and are discussed in a later section.

In the B-S design, the mean score on the pretest for the POWERSOURCE© group was 17.64 and for the control group 17.36 (Table 6). This indicates that two groups in the B-S design are equivalent in their pretest score. However, in the within-schools (W-S) design, pretest scores for POWERSOURCE© students and control students were 19.01 and 17.62, respectively. Thus, in the W-S design the POWERSOURCE© students have a significantly higher pretest score than the students in the control group.

Analysis of the transfer measure indicated that in the B-S design, the POWERSOURCE© students scored a mean of 17.25, whereas the mean for the control students was 16.83 (Table 7). The observed mean difference is approximately 0.42, which is a 0.08 pooled standard deviation. In the W-S design, the POWERSOURCE© students had a higher mean score on the transfer measure than control students by approximately 1.7, which is the size of a 0.3 pooled-standard deviation.

Table 6

Descriptive Statistics of Pretest Scores

Design			Pretest total score			
Design	Control/Treatment	<i>N</i>	Mean	<i>SD</i>	Min	Max
Between	Control	633	17.36	4.79	5	28
	Treatment	842	17.64	4.84	1	27
Within	Control	1120	17.62	4.36	2	27
	Treatment	1496	19.01	4.24	1	28

Table 7

Descriptive Statistics of Transfer Measure Scores

Design			Transfer measure total			
Design	Control/Treatment	<i>N</i>	Mean	<i>SD</i>	Min	Max
Between	Control	633	16.83	5.51	2	32
	Treatment	842	17.25	6.37	1	33
Within	Control	1120	17.15	5.38	1	31
	Treatment	1496	18.93	6.12	0	33

As described earlier, the POWERSOURCE[®] intervention addresses three big ideas. Items on both the pretest and transfer measure assess student understanding of the concepts of rational number equivalence (RNE), properties of arithmetic (PA) and principles for solving equations (SE). Analyses were also completed on three different subscores related to these conceptual areas.

Properties of arithmetic (PA). There were 8 items addressing PA on the pretest, and 5 items on the transfer measure. PA items on the transfer measure seem to be more difficult than those on the pretest. On average, students got correct more than half the items in the pretest, but on the transfer measure performance, the PA items were low. Overall POWERSOURCE[®] students scored higher on both the pretest and transfer measure PA items than students in the control group (Tables 8 & 9). The mean score for the PA items on the transfer measure for the B-S POWERSOURCE[®] students was 1.65 and for the control group 1.21 (a difference of 0.44, which is about a 0.38 pooled-standard deviation). Furthermore, this difference is even larger, 0.77 (0.64 pooled-*SD*) in the W-S design with POWERSOURCE[®] students again scoring higher than the control group students.

Table 8

Descriptive Statistics of PA Scores on the Pretest

Design			PA_pretest			
Design	Control/Treatment	<i>N</i>	Mean	<i>SD</i>	Min	Max
Between	Control	633	4.72	1.96	0	8
	Treatment	842	4.80	1.99	0	8
Within	Control	1120	4.85	1.86	0	8
	Treatment	1496	5.35	1.77	0	8

Table 9

Descriptive Statistics of PA scores the Transfer Measure

Design			PA_posttest			
Design	Control/Treatment	<i>N</i>	Mean	<i>SD</i>	Min	Max
Between	Control	633	1.21	1.05	0	5
	Treatment	842	1.65	1.28	0	5
Within	Control	1120	1.12	1.00	0	5
	Treatment	1496	1.89	1.41	0	5

Rational Number Equivalence (RNE). There were 6 items addressing RNE on the pretest, and 11 items on the transfer measure. Similar to PA, the mean pretest score for the POWERSOURCE[®] students in the B-S design on the RNE items was higher ($M = 3.82$) than for the control students ($M = 3.73$), with a difference between the two groups of 0.09 (Table 10). Mean scores on the transfer measure were similar between the POWERSOURCE[®] ($M = 6.29$) and the control group ($M = 6.03$) in the B-S design. In the W-S design, however, POWERSOURCE[®] students scored higher ($M = 6.79$) on the RNE items than did the control group ($M = 6.29$) by approximately 0.23 pooled standard deviation (Table 11).

Table 10

Descriptive Statistics of RNE Scores on the Pretest

Design			RNE Pretest Score			
Design	Control/Treatment	<i>N</i>	Mean	<i>SD</i>	Min	Max
Between	Control	633	3.73	1.41	0	6
	Treatment	842	3.82	1.32	0	6
Within	Control	1120	3.72	1.33	0	6
	Treatment	1496	4.06	1.23	0	6

Table 11

Descriptive Statistics of RNE Scores on the Transfer Measure

Design			RNE Transfer Measure Score			
Design	Control/Treatment	<i>N</i>	Mean	<i>SD</i>	Min	Max
Between	Control	633	6.03	2.24	0	10
	Treatment	842	6.29	2.34	0	11
Within	Control	1120	6.29	2.23	0	10
	Treatment	1496	6.79	2.18	0	11

Solving Equations (SE). We included 7 SE items on the pretest and 14 on the transfer measure. The pretest scores (Table 12) were very similar across the four different groups (control and POWERSOURCE[®] groups in the between- and within-school designs). On the transfer measure, POWERSOURCE[®] students in the W-S design scored higher than the control students by 0.55, whereas control students outperformed POWERSOURCE[®] students in the B-S design by 0.16 point (Table 13).

Table 12

Descriptive Statistics of SE Scores on the Pretest

Design			SE_pretest			
Design	Control/Treatment	<i>N</i>	Mean	<i>SD</i>	Min	Max
Between	Control	633	4.53	1.37	1	7
	Treatment	842	4.52	1.41	0	7
Within	Control	1120	4.55	1.24	0	7
	Treatment	1496	4.89	1.21	0	7

Table 13

Descriptive Statistics of SE Scores on the Transfer Measure

Design			SE_posttest			
Design	Control/Treatment	<i>N</i>	Mean	<i>SD</i>	Min	Max
Between	Control	633	7.34	2.81	0	14
	Treatment	842	7.18	3.14	0	14
Within	Control	1120	7.54	2.81	0	14
	Treatment	1496	8.09	2.97	0	14

Missing Data

It is very common to have missing data either in pretest or posttest measures in randomized cluster trials. Attrition from a sample can undermine the validity of a study, causing groups to be non-comparable on factors unrelated to the intervention. An intent-to-treat analysis is used to preserve the effects of the randomization procedure—that is, “as randomized, as analyzed.” In this study, our unit of assignment is the teacher in W-S design and school in the B-S design. As mentioned above two schools were dropped out of this study and all the rest of schools and teachers remained in the study after randomization.

For missing student pretest and posttest data, we examined whether *missingness* can be linked to the intervention. We will also examined the association between the missingness and student background characteristics. For example, we calculated the mean score on the transfer measure for students who do not have missing data in their pretest versus those who have missing data. Likewise, we calculated the mean of pretest of students who do not have missing data in their posttest versus those who have missing data. These results provide important information about whether missingness is systematically related to intervention or student characteristics.

As can be seen in Table 14, the percentage of complete cases was 72.6%. The treatment group had 74.4 % of complete cases, whereas the control group had 71.2%. Taking treatment into consideration, the rate of missingness in the transfer measure scores for the control group (24.0%) was higher than that in the POWERSOURCE[®] group (13.6%). In contrast, the rate of missingness in pretest scores was higher in the POWERSOURCE[®] group (11.3%) than in control group (5.9%). With respect to design, the missing rate in pretest scores was similar to each design—B-S design (8.3%) compared to the W-S design (9.0%).

Table 14

Missing Data Patterns in Pretest and Transfer Measure (2007–2008 School Year)

Data pattern	Between-school (B-S) design				Within-school (W-S) design			
	Control		Treatment		Control		Treatment	
	Freq.	%	Freq.	%	Freq.	%	Freq.	%
No missing	633	64.1	842	77.2	1120	76.0	1496	72.9
Missing in pretest	67	6.8	106	9.7	74	5.0	264	12.9
Missing in transfer measure	288	29.1	142	13.0	279	18.9	292	14.2
Total	988	100.0	1090	100.0	1473	100.0	2052	100.0

As such, a fairly high number of students (approximately 27%) have missing data either for the pretest or the transfer measure. Among the four different groups, control group students in the B-S design have a distinctly higher percentage of missing data on the transfer measure (29.1%, compared to the rest of groups'; on average, 15.4% missingness on the transfer measure). However, this group has a very small percentage of missing pretest data. In general, there are more missing data for the transfer measure than for the pretest, and this kind of pattern is well expected, partly because of high rate of student mobility in large urban schools.

To determine if there is any possible association between missing data and test scores, descriptive statistics for each available case were examined (see Tables 15 and 16). First, the mean pretest scores for students with complete data was higher ($mean = 18.1$) than the mean for the group of students who were missing transfer measure scores ($mean = 16.8$), indicating that the students whose transfer measure scores are missing could have lower pretest scores. Similarly, the mean transfer measure score for those students with complete data was higher ($mean = 17.8$) than the group of students who do not have pretest scores ($mean = 15.8$), indicating that the students whose pretest scores are missing could have lower transfer measure scores.

The key concern in this kind of pattern is whether the missing data favors one of the two groups. For example, let us consider the case of data analysis only with complete cases, (i.e., excluding cases which have missingness either in pretest or transfer measure). As a result of this kind of “list-wise deletion,” (i.e., if students in the control group who have

higher transfer measure score are excluded due to the missingness of their pretest scores), comparison between control and treatment group is biased to the treatment group. In contrast, if treatment students who have higher transfer measure scores are excluded due to the missingness of their pretest scores, the comparison is unfavorable to treatment group.

Taking this into account, there is no “unfavorable” missingness for control or treatment groups either in pretest or in posttest. As can be seen in Table 14, among students who do have a transfer measure score, students in the POWERSOURCE[®] group have slightly higher pretest scores ($M = 18.5$) than those in the control group ($M = 17.5$). Similarly, the mean pretest scores for the students whose posttest scores are missing ($M = 16.8$) was about the same in both POWERSOURCE[®] and control groups. However, the mean transfer measure score for the POWERSOURCE[®] students whose pretest scores are missing, was slightly higher (16.1) than those in control group (15.2). Thus, analysis only with complete cases might underestimate the treatment effects. In other words, the complete cases analysis might yield a conservative estimate of the POWERSOURCE[®] effect.

We analyzed the two data sets and compared the results as a sensitivity analysis. First, we fit a 2-level hierarchical model as presented in the following statistical model section. Second, we imputed the missing data both in the pretest and transfer measure as described below and fit the same 2-level hierarchical model as used for complete cases. Third, we examined whether the results are consistent to each other. As to imputation of missing values, we assume that transfer measure score distribution is normal, conditional on pretest scores as well as vice versa. The conditional mean substitution using four major stratifications was made to impute missing values. Therefore, samples were stratified by four: between-control, between-treatment, within-control, and within-treatment, and missing values in the transfer measure were imputed by the mean of available posttest scores of the students who had the same pretest scores as the missing students. Similarly, missing values in pretest scores were imputed by the mean of available pretest scores of the students who have the same transfer measure scores as the student whose pretest score is missing. We did this even though substituting the means results in an underestimation of variance.

Table 15

Descriptive Statistics of Pretest Scores by Missing Pattern

Data pattern	Between design						Within design					
	Control			Treatment			Control			Treatment		
	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>
No missing	633	17.4	4.8	842	17.6	4.8	1120	17.6	4.4	1496	19.0	4.2
Missing in posttest	288	16.7	5.0	142	17.1	5.7	279	16.9	4.5	292	16.6	4.9

Table 16

Descriptive Statistics of Transfer Measure Scores by Missing Pattern

Data pattern	Between design						Within design					
	Control			Treatment			Control			Treatment		
	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>
No missing	633	16.8	5.5	842	17.3	6.4	1120	17.1	5.4	1496	18.9	6.1
Missing in pretest	67	15.6	5.0	106	15.2	6.4	74	14.8	6.0	264	16.5	5.7

Statistical Model: 2-Level Hierarchical Model

As described earlier, we used two different designs (between- and within-schools) due to districts needs and configuration. Incorporating two types of randomization within the same study could be a common problem in a large-scale cluster randomized study where whole schools participate as control sites and receive no treatment. Given the difficulty recruiting school sites, it was necessary for us to be flexible in our design plan and allow schools to implement the type of design they found the most appealing.

The sample-size in the B-S design (36 teachers, 15 schools) and the W-S design (49 teachers and 10 schools) was fairly small. As such, the statistical power of the key parameter of interest, (i.e., treatment effect), is not as high we would like. Furthermore, if data were analyzed separately—which would be the easiest approach—we would have to synthesize the results and thus lose more statistical power. Lastly there is a concern on the choice of unit

of analysis. At first glance, the data appears to have a 3-level hierarchical structure (students are nested within teachers who in turn are nested within schools) in both designs. However, given the very small number of teachers in some schools, especially in the B-S design where most of schools only have two or three teachers, it does not seem appropriate to use “teacher” as another level in our hierarchical model. One possible solution for both designs is to use a 2-level hierarchical model, (i.e., students in level-1 and schools in level-2). However, this method ignores W-S individual teacher variability. Furthermore, all the valuable teacher information (e.g., three different individual teacher pre- and post-surveys) can be used only as a school aggregate.

Taking the above methodological concerns into account, we used a 2-level hierarchical model (HM) to examine the POWERSOURCE[®] effects on the transfer measure outcome. In order to synthesize two different designs and address the unit of analysis issue, we chose “teacher” as a unit of analysis with individual school effects in the model. School specific fixed effects address school blocking factors and intra-class correlation of school. As such, we can examine whether there is a differential treatment effect depending upon two different designs not at the cost of losing statistical power. The level-1, between-student; within-teacher, model, specifies the relationship between student score on the transfer measure and his or her pretest score as a covariate. The transfer measure total score, Y_{ij} , is the outcome for student i in teacher j . The pretest score for student i in teacher j ($Pretest_{ij}$) is centered around its mean. By virtue of this centering method, β_{0j} is the unadjusted transfer measure mean for teacher j , and β_{1j} is the pretest-outcome slope for teacher j .

Level-1 (between-student; within-teacher) model:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(Pretest_{ij} - Pretest_{.j}) + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, \sigma^2) \quad (1)$$

The level 2 (between-teacher) model includes the treatment indicator variable (control group teacher = 0 and POWERSOURCE[®] teacher = 1), design indicator variable (W-S design = 0 and B-S design = 1), and pretest mean. Note that we also include school flag variables in order to estimate school specific effects, which takes into account intra-class correlation in school level.

Level-2 (between-teacher) model:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Trt_j + \gamma_{02}Design_j + \gamma_{03}Trt_j \times Design_j + \gamma_{04}(Pretest_{.j} - Pretest_{.}) + \gamma_{0_k}S_{_k} + \dots + \gamma_{0_kn-3}S_{_kn-3} + u_{0j} \quad u_{0j} \sim N(0, \tau_{00}) \quad (2a)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Trt_j + u_{1j} \quad u_{1j} \sim N(0, \tau_{11}) \quad (2b)$$

γ_{00} represents the expected mean for control group in the W-S design holding other variables constant including school specific effects. γ_{01} and γ_{02} are main effects of the treatment and design, respectively. γ_{03} captures the interaction effect between treatment condition and design. If this coefficient is statistically significant, it indicates that the treatment effect is different depending upon designs. γ_{0_k} through $\gamma_{0_{kn-3}}$ are school specific fixed effects. Note that there are $k-3$ school fixed effects, where k is the total number of schools and there are four baseline groups: treatment and control in the B-S design, and treatment and control in the W-S design. γ_{11} captures the difference in the pretest-outcome slope between treatment and control group.

2-Level Hierarchical Model Results

Transfer measure (posttest) outcome (total score). We first present HM results where the total score on the transfer measure is used as outcome. Table 17 presents estimates of all the fixed effect parameters and variance components in the model. As seen in this table, there is no statistically significant main effect either in treatment or in design. In addition, the treatment and design interaction effect is not statistically significant. These results indicate that treatment effect does not differ by two different designs. However, we found a main effect of the pretest score. The estimate of the pretest mean is 1.18 and its p -value is smaller than .0001. The students in classes with higher pretest mean scores tend to have higher posttest mean scores as well.

Table 17

HM result: Transfer Measure Total Score

Fixed effects	Coefficient	SE	df	t-value	p-value
Model for class mean					
Intercept, γ_{00}	17.20	0.81	58	21.19	< .0001
Treatment, γ_{01}	0.40	0.55	58	0.71	0.479
Design, γ_{02}	1.50	1.95	58	0.77	0.444
Treatment*Design, γ_{03}	-2.12	2.03	58	-1.05	0.300
Pretest Mean, γ_{04}	1.18	0.10	58	12.08	< .0001
s01, γ_{0_01}	0.12	2.23	58	0.05	0.958
s02, γ_{0_02}	-0.30	2.01	58	-0.15	0.881
s03, γ_{0_03}	0.90	2.49	58	0.36	0.719
s04, γ_{0_04}	0.74	1.57	58	0.47	0.641
s05, γ_{0_05}	1.41	1.03	58	1.37	0.175
s06, γ_{0_06}	-2.54	1.06	58	-2.39	0.020
s07, γ_{0_07}	-3.02	2.51	58	-1.20	0.235
s08, γ_{0_08}	1.56	1.51	58	1.03	0.305
s09, γ_{0_09}	2.42	1.38	58	1.75	0.085
s10, γ_{0_10}	1.28	1.17	58	1.10	0.276
s11, γ_{0_11}	-1.92	2.05	58	-0.94	0.353
s13, γ_{0_13}	-0.49	2.16	58	-0.23	0.821
s14, γ_{0_14}	0.17	1.06	58	0.16	0.871
s16, γ_{0_16}	0.69	1.18	58	0.59	0.560
s17, γ_{0_17}	1.22	1.06	58	1.15	0.255
s18, γ_{0_18}	6.74	1.57	58	4.30	< .0001
s19, γ_{0_19}	-0.55	1.09	58	-0.50	0.617
s20, γ_{0_20}	-1.49	2.55	58	-0.58	0.561
s21, γ_{0_21}	0.80	1.06	58	0.76	0.453
s23, γ_{0_22}	0.91	1.50	58	0.61	0.544
s24, γ_{0_24}	-0.90	1.37	58	-0.66	0.512
s25, γ_{0_25}	-1.27	2.26	58	-0.56	0.576
Model for pretest slope					
Intercept, γ_{10}	0.52	0.05	4004	10.13	< .0001
Treatment, γ_{11}	0.24	0.07	4004	3.51	0.000

(table continues)

Table 17 (*continued*)

Random effects	Variance component	SE	z-value	p-value
Class mean, u_{0j}	2.87	0.63	4.55	< .0001
Pretest slope, u_{1j}	0.07	0.01	4.45	< .0001
Level-1 error for B-S, e_{ij_1}	18.91	0.71	26.57	< .0001
Level-1 error for W-S, e_{ij_2}	17.87	0.50	35.48	< .0001

Interestingly, we found a significant interaction effect between treatment and student pretest. For both W-S and B-S designs, the pretest-outcome slope is steeper for treatment group students than for the control group students: The estimate of the interaction coefficient is positive, 0.24 (p -value is 0.000). This indicates that students with higher pretest scores tend to benefit more from the treatment as compared to students with lower pretest scores. This tendency is explicitly shown in Figure 7. This figure presents model-based fitted relationships between pretest and posttest for four different groups. In both the B-S and the W-S designs, the two fitted lines are crossed. Specifically, in the B-S design, the two fitted lines are crossed at approximately 1.0 SD below of the pretest score mean, whereas in the W-S design, the two fitted lines are crossed at 0.5 SD below the pretest score mean. This figure shows that POWERSOURCE[®] students whose pretest scores are higher than the pretest mean score have higher posttest scores than those in control group. In other words, students with higher pretest scores tend to benefit appreciably more from the POWERSOURCE[®] treatment than students with middle or lower pretest scores. Students whose pretest score is 0.5 standard deviation or more above the pretest mean, perform about 1.3 points higher on the posttest, approximately 1/4 of a standard deviation. Furthermore, students with higher pretest scores (at 2 SD of pretest score mean) perform about 3 points higher on the posttest, slightly more than 1/2 of a standard deviation. The W-S design had a similar pattern as the B-S design. In the W-S design, POWERSOURCE[®] students had higher posttest scores than the control students from approximately 0.5 SD below the pretest mean. At the higher end of the pretest score the difference is about 2.5 points, approximately 0.4 pooled standard deviation.

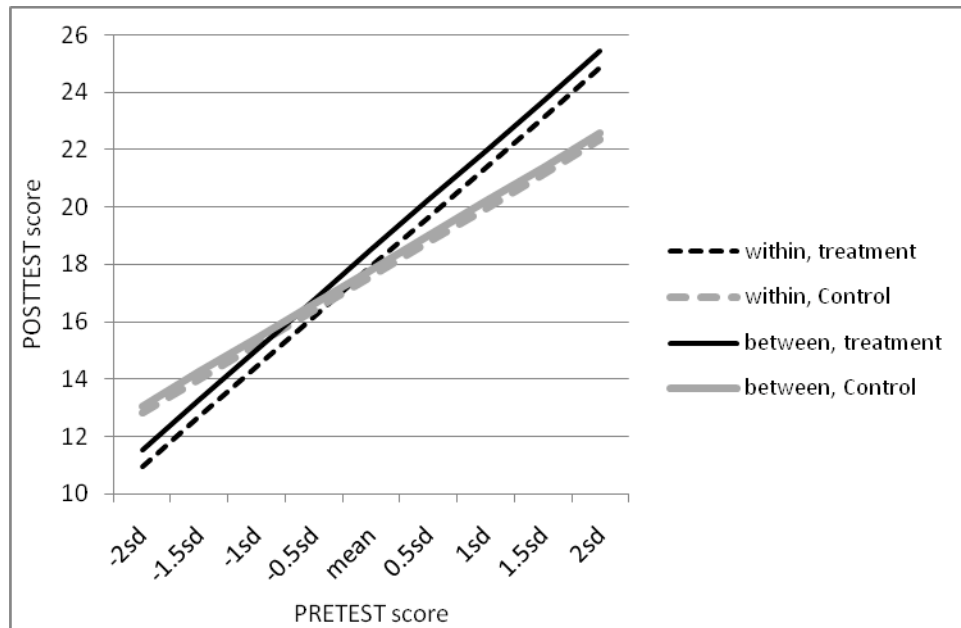


Figure 7. HM result (transfer measure total score): Fitted relationships between pretest score and posttest by design and treatment condition.

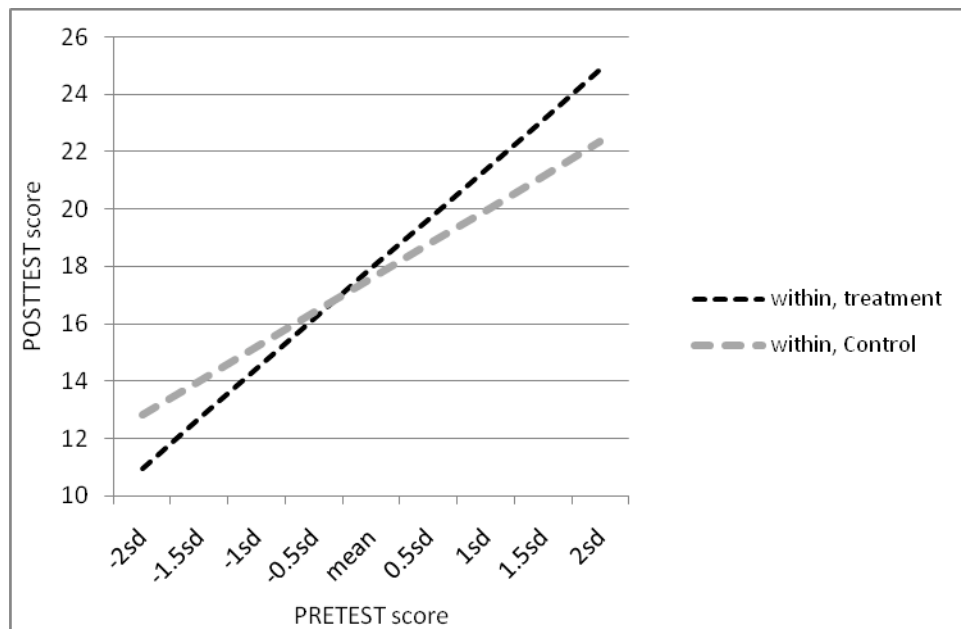


Figure 8(a). HM result: Fitted relationship between pretest and posttest for treatment conditions in within-school design.

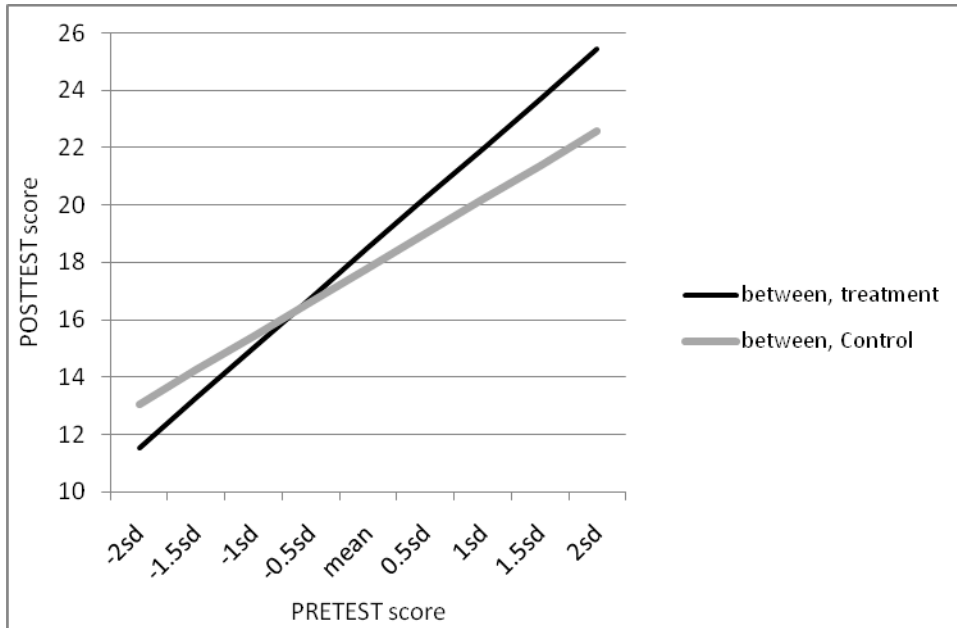


Figure 8(b). HM result: Fitted relationship between pretest and posttest for treatment conditions in between-school design.

HM result: Transfer measure subscore (RNE). The transfer measure contained items relating to all POWERSOURCE domains. We fit the same 2-level HM as presented in Equations 1, 2a, and 2b to the items addressing RNE concepts (11 items). As can be seen in Table 18, there was no significant treatment or design effect. In addition, neither the interaction effect of treatment and design or of treatment and student pretest score was statistically significant. Figure 9 presents the fitted relationship between RNE pretest subscore and RNE posttest subscore by design and treatment condition.

Table 18

HM result: RNE Transfer Measure Subscore

Fixed effects	Coefficient	SE	df	t-value	p-value
Model for class mean					
Intercept, γ_{00}	6.09	0.31	58	19.86	< .0001
Treatment, γ_{01}	-0.06	0.22	58	-0.30	0.767
Design, γ_{02}	0.24	0.73	58	0.33	0.740
Treatment*Design, γ_{03}	-0.27	0.77	58	-0.34	0.732
Pretest Mean, γ_{04}	1.77	0.20	58	9.00	< .0001
s01, γ_{0_01}	0.01	0.84	58	0.01	0.991
s02, γ_{0_02}	0.21	0.75	58	0.29	0.776
s03, γ_{0_03}	-0.17	0.92	58	-0.18	0.855
s04, γ_{0_04}	0.73	0.61	58	1.20	0.236
s05, γ_{0_05}	0.61	0.39	58	1.58	0.121
s06, γ_{0_06}	-1.03	0.40	58	-2.57	0.013
s07, γ_{0_07}	-0.91	0.94	58	-0.97	0.337
s08, γ_{0_08}	0.90	0.57	58	1.57	0.121
s09, γ_{0_09}	0.42	0.55	58	0.78	0.441
s10, γ_{0_10}	0.89	0.43	58	2.07	0.043
s11, γ_{0_11}	-0.56	0.77	58	-0.72	0.474
s13, γ_{0_13}	0.53	0.81	58	0.66	0.513
s14, γ_{0_14}	0.53	0.40	58	1.32	0.191
s16, γ_{0_16}	0.43	0.45	58	0.95	0.344
s17, γ_{0_17}	1.02	0.40	58	2.53	0.014
s18, γ_{0_18}	2.56	0.61	58	4.21	< .0001
s19, γ_{0_19}	0.23	0.41	58	0.56	0.575
s20, γ_{0_20}	0.40	0.97	58	0.41	0.680
s21, γ_{0_21}	1.03	0.39	58	2.62	0.011
s23, γ_{0_22}	0.83	0.56	58	1.46	0.148
s24, γ_{0_24}	0.05	0.53	58	0.09	0.929
s25, γ_{0_25}	0.94	0.86	58	1.09	0.279
Model for pretest slope					
Intercept, γ_{10}	0.42	0.05	4004	8.65	< .0001
Treatment, γ_{11}	0.10	0.06	4004	1.50	0.135

(table continues)

Table 18 (continued)

Random effects	Variance component	SE	z-value	p-value
Class mean, u_{0j}	0.41	0.09	4.43	< .0001
Pretest slope, u_{1j}	0.03	0.01	2.37	0.009
Level-1 error for B-S, e_{ij_1}	4.03	0.15	26.73	< .0001
Level-1 error for W-S, e_{ij_2}	3.58	0.10	35.47	< .0001

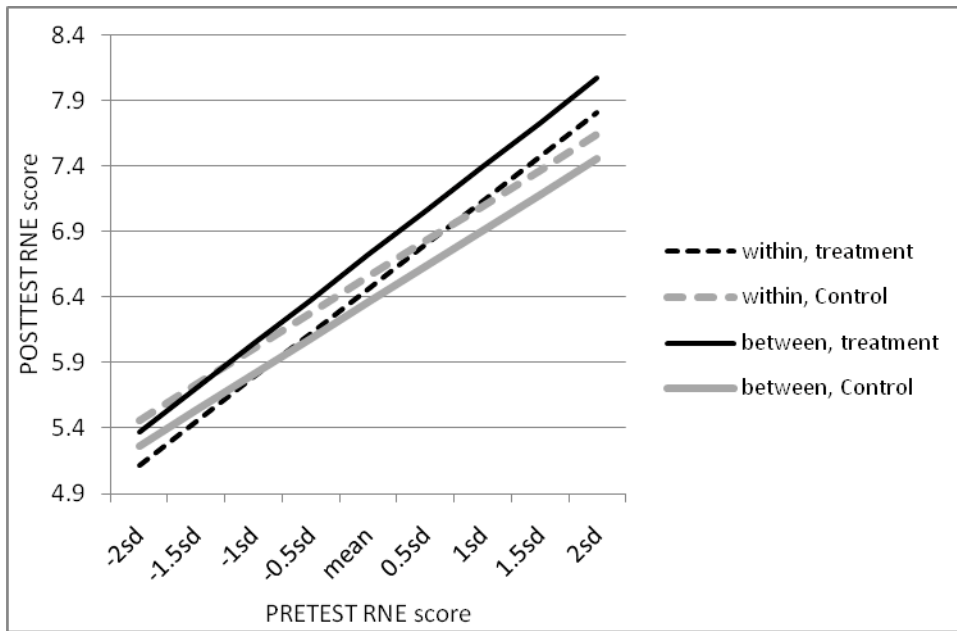


Figure 9. HM result (RNE transfer measure subscore): Fitted relationships between pretest RNE subscore and posttest RNE subscore by design and treatment condition.

HM result: Transfer measure subscore (PA). Figure 10 presents the results from items assessing knowledge of properties of arithmetic (specifically the distributive property) of which there were five on the transfer measure. As can be seen in Table 19, on these items we found that the POWERSOURCE[®] effect was statistically significant (estimate = 0.57, p -value = 0.002) and the interaction effect between pretest and treatment condition was also significant (estimate = 0.13, p -value = 0.000). These results clearly suggest that for properties of arithmetic, POWERSOURCE[®] students both in W-S and B-S designs significantly outperformed students in control groups, and this effect gets larger for those whose pretest scores are higher (see Figure 10).

Table 19

HM result: PA transfer measure subscore

Fixed effects	Coefficient	SE	df	t-value	p-value
Model for class mean					
Intercept, γ_{00}	1.13	0.25	58	4.44	< .0001
Treatment, γ_{01}	0.57	0.17	58	3.28	0.002
Design, γ_{02}	0.11	0.61	58	0.17	0.862
Treatment*Design, γ_{03}	-0.04	0.63	58	-0.06	0.955
Pretest Mean, γ_{04}	0.41	0.08	58	5.11	< .0001
s01, γ_{0_01}	0.04	0.69	58	0.06	0.956
s02, γ_{0_02}	-0.09	0.63	58	-0.14	0.893
s03, γ_{0_03}	0.47	0.78	58	0.61	0.544
s04, γ_{0_04}	-0.35	0.48	58	-0.73	0.467
s05, γ_{0_05}	0.17	0.32	58	0.53	0.595
s06, γ_{0_06}	-0.23	0.33	58	-0.68	0.496
s07, γ_{0_07}	-0.05	0.78	58	-0.07	0.945
s08, γ_{0_08}	0.14	0.47	58	0.30	0.765
s09, γ_{0_09}	0.20	0.42	58	0.48	0.634
s10, γ_{0_10}	-0.04	0.36	58	-0.10	0.922
s11, γ_{0_11}	0.11	0.64	58	0.18	0.861
s13, γ_{0_13}	-0.04	0.67	58	-0.06	0.955
s14, γ_{0_14}	0.27	0.33	58	0.80	0.426
s16, γ_{0_16}	0.27	0.36	58	0.74	0.459
s17, γ_{0_17}	0.27	0.33	58	0.80	0.428
s18, γ_{0_18}	0.54	0.48	58	1.13	0.263
s19, γ_{0_19}	-0.09	0.34	58	-0.26	0.797
s20, γ_{0_20}	-0.50	0.79	58	-0.63	0.533
s21, γ_{0_21}	0.21	0.33	58	0.65	0.520
s23, γ_{0_22}	-0.31	0.46	58	-0.67	0.504
s24, γ_{0_24}	-0.23	0.42	58	-0.56	0.575
s25, γ_{0_25}	0.07	0.69	58	0.10	0.918
Model for pretest slope					
Intercept, γ_{10}	0.04	0.02	4004	2.56	0.011
Treatment, γ_{11}	0.13	0.02	4004	5.74	< .0001

(table continues)

Table 19 (continued)

Random effects	Variance component	SE	z-value	p-value
Class mean, u_{0j}	0.30	0.06	4.73	< .0001
Pretest slope, u_{1j}	0.00	0.00	1.67	0.047
Level-1 error for B-S, e_{ij_1}	1.10	0.04	26.61	< .0001
Level-1 error for W-S, e_{ij_2}	1.16	0.03	35.61	< .0001

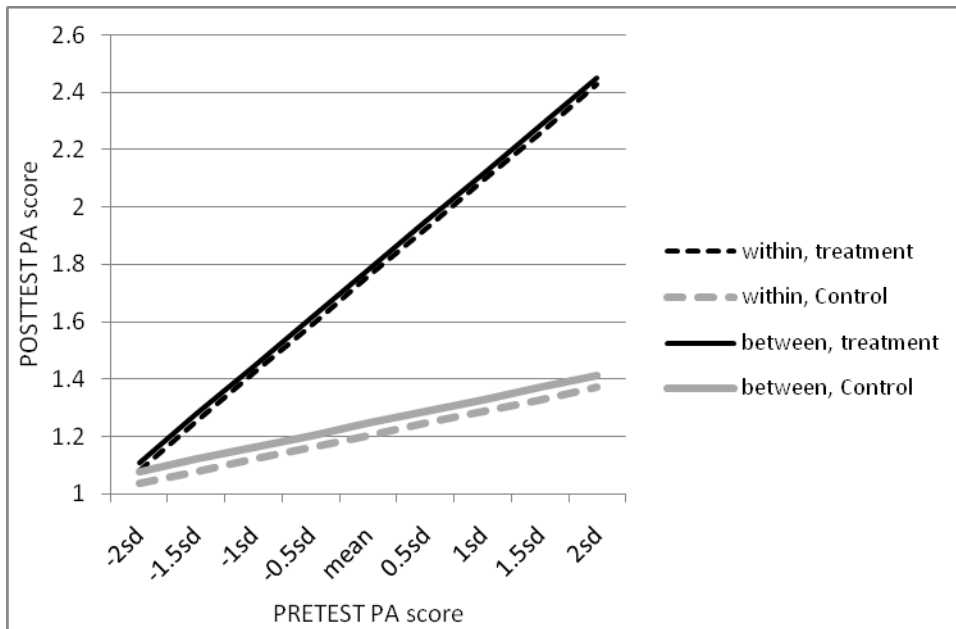


Figure 10. HM result (PA transfer measure subscore): Fitted relationships between pretest PA subscore and posttest PA subscore by design and treatment condition.

HM result: transfer measure subscore (SE). The transfer measure contained 14 items related to solving equations (SE). Table 20 and Figure 11 present the HM results for the SE items. As we found in RNE, there were no statistically significant main effects of the treatment. However, the interaction effect between treatment and student pretest is approaching statistical significance (estimate = 0.2, p -value = 0.074). This suggests that the treatment effect may be greater for those students with a higher initial pretest score. Note that the interaction effect between treatment and design is also approaching statistical significance (estimate = -1.81, p -value = 0.068). This suggests that the treatment effect may differ by design, that is, there is a larger treatment effect in the B-S design than in the W-S design.

Table 20

HM result: SE Transfer Measure Subscore

Fixed effects	Coefficient	SE	df	t-value	p-value
Model for class mean					
Intercept, γ_{00}	8.09	0.40	58	20.33	< .0001
Treatment, γ_{01}	0.09	0.27	58	0.34	0.739
Design, γ_{02}	0.45	0.94	58	0.48	0.631
Treatment*Design, γ_{03}	-1.81	0.97	58	-1.86	0.068
Pretest Mean, γ_{04}	2.07	0.19	58	11.03	< .0001
s01, γ_{0_01}	-0.11	1.08	58	-0.10	0.919
s02, γ_{0_02}	-1.44	0.96	58	-1.50	0.140
s03, γ_{0_03}	0.45	1.20	58	0.38	0.707
s04, γ_{0_04}	1.63	0.77	58	2.13	0.038
s05, γ_{0_05}	0.44	0.50	58	0.88	0.385
s06, γ_{0_06}	-1.85	0.52	58	-3.54	0.001
s07, γ_{0_07}	-2.30	1.21	58	-1.89	0.063
s08, γ_{0_08}	0.22	0.74	58	0.29	0.772
s09, γ_{0_09}	1.56	0.68	58	2.29	0.026
s10, γ_{0_10}	0.54	0.56	58	0.95	0.347
s11, γ_{0_11}	-1.67	0.99	58	-1.70	0.095
s13, γ_{0_13}	-0.91	1.04	58	-0.88	0.385
s14, γ_{0_14}	-1.22	0.52	58	-2.33	0.023
s16, γ_{0_16}	0.34	0.58	58	0.59	0.557
s17, γ_{0_17}	-0.16	0.52	58	-0.31	0.756
s18, γ_{0_18}	3.75	0.79	58	4.77	< .0001
s19, γ_{0_19}	-1.21	0.53	58	-2.29	0.026
s20, γ_{0_20}	-0.93	1.24	58	-0.75	0.455
s21, γ_{0_21}	-0.69	0.52	58	-1.32	0.191
s23, γ_{0_22}	0.78	0.73	58	1.07	0.290
s24, γ_{0_24}	-0.11	0.66	58	-0.17	0.866
s25, γ_{0_25}	-0.62	1.10	58	-0.56	0.575
Model for pretest slope					
Intercept, γ_{10}	0.67	0.09	4004	7.89	< .0001
Treatment, γ_{11}	0.20	0.11	4004	1.78	0.074

(table continues)

Table 20 (continued)

Random effects	Variance component	SE	z-value	p-value
Class mean, u_{0j}	0.64	0.15	4.34	< .0001
Pretest slope, u_{1j}	0.15	0.04	3.8	< .0001
Level-1 error for B-S, e_{ij_1}	6.15	0.23	26.67	< .0001
Level-1 error for W-S, e_{ij_2}	5.73	0.16	35.42	< .0001

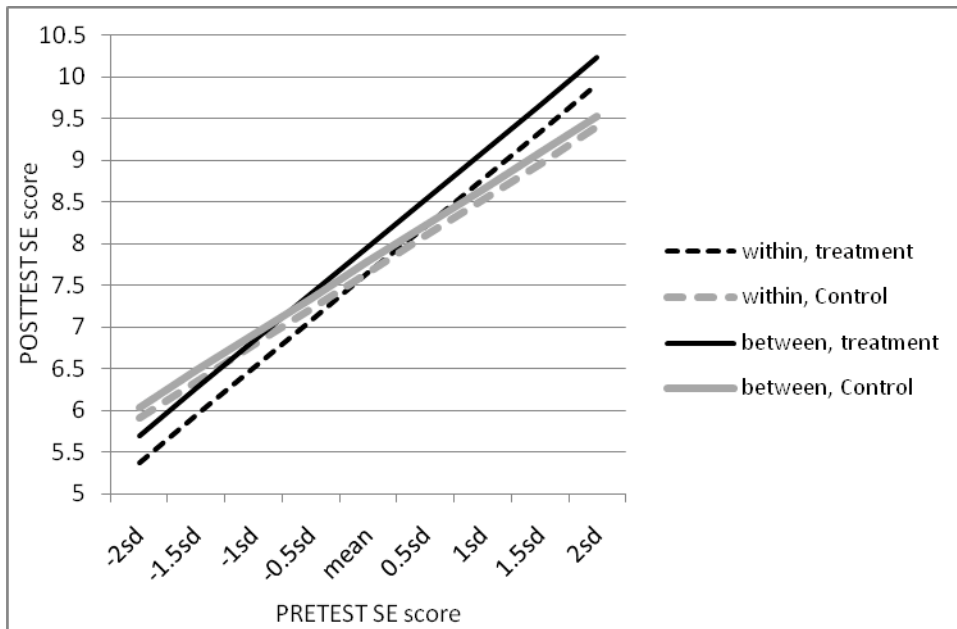


Figure 11. HM result (SE transfer measure subscore): Fitted relationships between pretest SE subscore and posttest SE subscore by design and treatment condition.

Sensitivity analysis: HM results with transfer measure total score with missing values imputed by the conditional mean scores. Table 21 represents the HM result based on the transfer measure scores with missing values imputed by the conditional mean scores. Although there are slight changes in the magnitude of coefficients compared to the ones based on the complete data, the general patterns are very similar to each of the original results. It is noticeable, however, that the p -values of the design main effect coefficient, and of design and treatment interaction coefficient become much smaller.

Table 21

HM Result (Transfer Measure Total Score, Missing Values Imputed by the Conditional Mean Scores)

Fixed effects	Coefficient	SE	df	t-value	p-value
Model for class mean					
Intercept, γ_{00}	16.95	0.62	71	27.38	< .0001
Treatment, γ_{01}	0.37	0.46	71	0.81	0.420
Design, γ_{02}	1.81	0.98	71	1.85	0.069
Treatment*Design, γ_{03}	-1.87	1.06	71	-1.76	0.083
Pretest Mean, γ_{04}	1.23	0.08	71	14.92	<.0001
s01, γ_{0_01}	-0.34	1.22	71	-0.28	0.783
s02, γ_{0_02}	-0.51	1.67	71	-0.31	0.760
s03, γ_{0_03}	-0.06	1.66	71	-0.03	0.972
s04, γ_{0_04}	-0.19	1.20	5820	-0.16	0.872
s05, γ_{0_05}	1.31	0.80	71	1.63	0.107
s06, γ_{0_06}	-2.30	0.86	71	-2.68	0.009
s07, γ_{0_07}	-2.35	1.16	71	-2.03	0.046
s08, γ_{0_08}	1.25	1.26	71	0.99	0.324
s09, γ_{0_09}	1.82	1.09	71	1.67	0.099
s10, γ_{0_10}	1.01	0.95	71	1.05	0.296
s11, γ_{0_11}	-1.62	1.09	71	-1.49	0.140
s13, γ_{0_13}	-0.97	1.29	71	-0.75	0.454
s14, γ_{0_14}	0.18	0.86	71	0.20	0.839
s16, γ_{0_16}	-0.14	0.91	71	-0.15	0.881
s17, γ_{0_17}	1.14	0.86	71	1.33	0.188
s18, γ_{0_18}	5.31	1.27	71	4.17	< .0001
s19, γ_{0_19}	-0.57	0.89	71	-0.64	0.524
s20, γ_{0_20}	-2.02	1.73	71	-1.17	0.247
s21, γ_{0_21}	0.46	0.85	71	0.54	0.591
s23, γ_{0_22}	0.52	1.17	71	0.44	0.662
s24, γ_{0_24}	-1.54	1.08	71	-1.42	0.160
s25, γ_{0_25}	-1.45	1.36	5820	-1.06	0.289
Model for pretest slope					
Intercept, γ_{10}	0.62	0.05	5820	13.03	< .0001
Treatment, γ_{11}	0.21	0.07	5820	3.27	0.001

(table continues)

Table 21 (continued)

Random effects	Variance component	SE	z-value	p-value
Class mean, u_{0j}	2.12	0.42	5.04	< .0001
Pretest slope, u_{1j}	0.08	0.02	5.1	< .0001
Level-1 error for B-S, e_{ij_1}	13.68	0.40	33.84	< .0001
Level-1 error for W-S, e_{ij_2}	15.40	0.37	41.42	< .0001

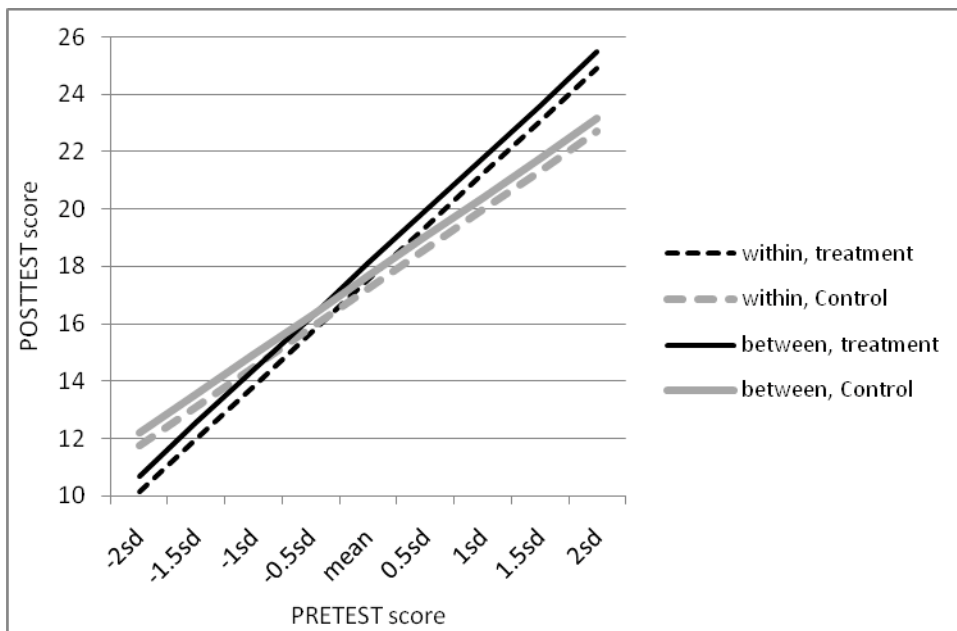


Figure 12. HM result (Transfer measure total score, missing values imputed by conditional mean scores): Fitted relationships between pretest and posttest by design and treatment condition.

Discussion

Results from the randomized study of the POWERSOURCE[®] formative assessment intervention yielded several interesting findings. First and foremost we see that a short amount of targeted intervention on key mathematical principles does have some impact on student performance on a transfer measure of related content. The POWERSOURCE[®] plan, includes approximately eight class periods of intervention in the classroom (instruction and assessment), with an additional 9 hours of professional development for teachers. Thus the period of intervention was very brief. POWERSOURCE[®] students, on average, did not outperform those in control groups, given that we did not find statistically significant main effects of the treatment. What we did find, however, was a significant interaction between

treatment and pretest scores. This indicates that students with higher scores on the pretest tend to benefit more from the POWERSOURCE[®] intervention compared to students with lower pretest scores. The effect size for those students is as high as a 0.5 pooled standard deviation. In other words, the POWERSOURCE[®] intervention had more impact on the higher-performing students than the lower-performing students.

Results of analyses of the transfer measure items related to the properties of arithmetic (specifically the distributive property) were the only ones where we saw a significant POWERSOURCE[®] effect. In both designs, students in the POWERSOURCE[®] group significantly outperformed control group students and the effect was larger as pretest scores increased. Item analyses indicated that PA items were more difficult for students than items focused on the other domains. In addition, qualitative data gathered from teacher interviews and observations revealed that teachers found their students to have a more difficult time with the PA content than the other POWERSOURCE[®] domains. Thus, we see a greater impact of POWERSOURCE[®] on the more difficult items.

As discussed above, we implemented a mixed-design plan based on district needs and configuration. For some districts with strong grade-level team collaborative initiatives, a W-S design seemed less tenable, as teachers would be collaborating on curricular issues creating potential for exposure of control group teachers to the POWERSOURCE[®] concepts and materials. Results of our analyses revealed no main effect of design indicating that it did not matter—in terms of looking at the magnitude of the POWERSOURCE[®] effect—which type of design (W-S or B-S) we used.

By design, the POWERSOURCE[®] instructional modules, assessments and professional development sessions complemented existing curricula and fit around what districts and schools already have in place. Our intent was to implement an intervention that would augment—not replace—mathematics instruction already in place in the districts. Time for POWERSOURCE[®] had to be found within sometimes tight curriculum frameworks and timelines. Our results, while promising, do not go as far as we would like in terms of improving student understanding of key big ideas and related skills in pre-algebra. It may be that the time allotted to our intervention is not sufficient to effect change in lower-performing students, even though we do see an impact with higher-performing students. Other measures—such as teacher implementation, teacher knowledge, and observations—may shed more light on the impact of POWERSOURCE[®] and how it is being implemented in the participating districts.

Given this is the second year of the POWERSOURCE[®] large-scale implementation, we are keenly interested in investigating differential/cumulative effects of the POWERSOURCE[®] experience both in students and teachers. As such, we will be closely analyzing teacher variables as we expect there to be a significant impact of number of years a teacher has been involved in POWERSOURCE.[®] We hope to see that teachers will become more proficient in their subject matter knowledge, more skilled in their formative use of assessment, and better equipped to focus their instruction on key ideas as they participate in the study. And, as a result, teachers will be more effective in helping students to improve their understanding of key algebra principles.

In the next year of the study we will examine the student growth trajectory with three time-series measures: a pretest (described above), an interim transfer measure (given midway through the year), and a post-transfer measure. We will work to address issues with missing data and increase our efforts to ensure all data are returned to us in a timely manner. Subsequent analyses will not only examine teacher and fidelity and implementation variables, but also student growth in performance on the Checks for Understanding assessments given to the POWERSOURCE[®] group. We are in the midst of conducting an equating study, where we will place scores from the Checks for Understanding in four different domains on a common scale. We will examine growth trajectories both within domain and across domains. The following questions will serve as the basis of our subsequent study: (a) What does the student growth trajectory look like across the year? (b) How much variability in student growth trajectory is observed?

We have obtained positive evidence on the instructional sensitivity of the tasks in this experimental study, and we will continue to study the value of the Checks for Understanding as formative assessments in the POWERSOURCE[®] program.

REFERENCES

- Baker, E. L. (1997). Model-based performance assessment. *Theory Into Practice*, 36(4), 247–254.
- Baker, E. L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 131–153). Englewood Cliffs, NJ: Prentice-Hall.
- Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice*, 5(1), 7–74.
- Brown, R. S., & Niemi, D. N. (2007). Investigating alignment of high school and community college assessments in California. San Jose, CA: National Center for Public Policy in Higher Education.
- California Department of Education. (2008). California Standardized Testing and Reporting (STAR). Sacramento, CA: Author. Retrieved from <http://star.cde.ca.gov/>
- Dimitrov, D. M. (2003). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement*, 27, 440–458.
- Herman, J. L., & Baker, E. L. (2006). Making benchmark testing work for accountability and improvement: Quality matters. *Educational Leadership*, 63(3), 48–55.
- Herman, J. L., & Gribbons, B. (2001). Lessons learned in using data to support school inquiry and continuous improvement: Final report to the Stuart Foundation (CSE Tech. Rep. No. 535). Los Angeles: University of California, Center for the Study of Evaluation (CSE), National Center for Research on Evaluation, Standards and Student Testing (CRESST).
- Herman, J., Osmundson, E., Ayala, C., Schneider, S., & Timms, M. (2006). *The nature and impact of teachers' formative assessment practices* (CSE Tech. Rep. No. 703). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Paper prepared for the Annual Meeting of the American Educational Research Association, Montreal, Canada. (April, 2005)
- Heritage, M., & Yeagley, R., (2005). Data use and school improvement: Challenges and prospects. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement. The 104th yearbook of the national society for the study of education. Part 2*. Malden, MA and Oxford, England. Blackwell Publishing.
- Howard, K., Vendlinski, T., Hemberg, B., Niemi, D., N., & Phelan, J. (2008, March). *Using error patterns formatively*. Presentation at the Annual Meeting for the American Educational Research Association, New York, NY.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.

- Niemi, D. N. (1996). *Instructional influences on content area explanations and representational knowledge: Evidence for the construct validity of measures of principled understanding*. (CSE Tech. Rep. No. 403). Los Angeles: University of California, Center for the Study of Evaluation (CSE), National Center for Research on Evaluation, Standards and Student Testing (CRESST).
- Niemi, D. N., Baker, E. L., & Sylvester, R. (2007). Scaling up, scaling down: Seven years of performance assessment development in the nation's second largest school district. *Educational Assessment*, 12, 195–214.
- Phelan, J., Choi, K., Vendlinski, T., Baker, E. L., & Herman, J. L. (2009, June). *Summary of recent findings: POWERSOURCE® formative assessment project*. Poster presentation at the Fourth Annual IES Research Conference, Washington DC.
- Phelan, J., Kang, T., Niemi, D. N., Vendlinski, T., & Choi, K. (2009). Some aspects of the technical quality of formative assessments in middle school mathematics (CRESST Report 750). Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).
- Plake, B. S., & Impara, J. C. (1997). Teacher assessment literacy: What do teachers know about assessment? In G. Phye (Ed.), *Handbook of classroom assessment* (pp. 53–68). San Diego, CA: Academic Press.
- Shepard, L. A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *Handbook of research on teaching*, (4th ed., pp. 1066–1101). Washington, DC: American Educational Research Association.
- Stiggins, R. J. (2005, December). From formative assessment to assessment FOR learning: A path to success in standards-based schools. *Phi Delta Kappan*, 87(4), 324–328.
- Stiggins, R. J. (2004). New assessment beliefs for a new school mission. *Phi Delta Kappan*, 86(1), 22.
- William, D., Lee, C., Harrison, C., & Black, P. (2004, March). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education*, 11(1), 49–65.
- Wolf, D., Bixby, J., Glenn, J. III, & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of research in education* (Vol. 17, pp. 31–74). Washington, DC: American Educational Research Association.

APPENDIX A

Item Analysis Results of POWERSOURCE[®] pretest

Item	Domain	<i>p</i> -value		Polyserial correlation	Rasch difficulty (on EAF scale)		IRT reliability (test Reli. = .917)	Alpha = .80
		0	1	$r_{poly.}$	b	$SE(b)$	Item reliability	If deleted
PRE01	PA	0.30	0.70	0.645	-0.562	0.019	0.33	0.79
PRE02	PA	0.40	0.60	0.673	-0.238	0.017	0.33	0.79
PRE03	RNE	0.11	0.89	0.795	-1.452	0.026	0.28	0.80
PRE04	SE	0.01	0.99	0.434	-2.792	0.066	0.14	0.80
PRE05	RA	0.71	0.30	0.410	0.670	0.018	0.32	0.80
PRE06	RA	0.37	0.63	0.316	-0.356	0.018	0.33	0.81
PRE07	RA	0.12	0.88	0.645	-1.382	0.026	0.28	0.80
PRE08	RA	0.06	0.94	0.521	-1.909	0.035	0.23	0.80
PRE09	SE	0.14	0.87	0.771	-1.279	0.024	0.29	0.79
PRE10	RNE	0.28	0.72	0.583	-0.649	0.019	0.32	0.80
PRE11	RNE	0.42	0.58	0.408	-0.186	0.017	0.33	0.80
PRE12	RNE	0.25	0.75	0.499	-0.748	0.020	0.32	0.80
PRE13	PA	0.15	0.85	0.704	-1.219	0.024	0.29	0.79
PRE14	PA	0.29	0.71	0.629	-0.619	0.019	0.32	0.79
PRE15	PA	0.43	0.57	0.682	0.006	0.016	0.33	0.79
PRE16	PA	0.37	0.63	0.514	-0.347	0.018	0.33	0.80
PRE17	RNE	0.40	0.60	0.524	-0.555	0.017	0.33	0.80
PRE18	SE	0.17	0.84	0.547	-1.120	0.023	0.30	0.80
PRE19	RA	0.22	0.78	0.406	-0.857	0.020	0.31	0.80
PRE20	RA	0.48	0.52	0.636	0.665	0.017	0.34	0.79
PRE21	SE	0.66	0.34	0.550	0.531	0.018	0.33	0.80
PRE22	PA	0.32	0.68	0.489	-0.502	0.019	0.33	0.80
PRE23	PA	0.87	0.13	0.110	1.348	0.025	0.09	0.81
PRE24	RNE	0.75	0.25	0.202	0.843	0.020	0.32	0.81
PRE25	SE	0.22	0.78	0.671	-0.863	0.020	0.31	0.79
PRE26	RA	0.57	0.43	0.397	0.259	0.017	0.33	0.80
PRE27	SE	0.91	0.10	0.314	1.601	0.028	0.26	0.80
PRE28	SE	0.33	0.67	0.700	-1.167	0.017	0.33	0.79

Note. PA = properties of arithmetic, RNE = rational number equivalence, SE = principles for solving linear equations, RA = application of core principles in these domains to other critical areas of mathematics, such as geometry and probability.

APPENDIX B

Student ID # _____

Powersource

Pretest

Circle the best answer for each question.

1. Evaluate the following: $6(2 + 3) =$

PA-PT-1

- a) 12
- b) 15
- c) 30
- d) 36

2. Simplify $7(4 + 1)$:

PA-PT-5

- a) $7 + 4 + 1$
- b) 12
- c) $(35) \cdot (5)$
- d) 35

3. Solve: $1 \cdot 87 =$

RN-PT-6

- a) 88
- b) 89
- c) 86
- d) 87

4. Solve: $4 + \square = 10$

SE-PT-1

- a) 8
- b) 10
- c) 4
- d) 6

5. Find the area of this square:

RA-PT-2



- a) 24 cm^2
- b) 36 cm^2
- c) 48 cm^2
- d) 18 cm^2

6. What is the sum of the internal angles of a triangle?

RA-PT-3

- a) 360°
- b) 270°
- c) 180°
- d) 220°

7. Sam is 7, and Jack is twice as old. How old is Jack?

RA-PT-4

- a) $10 \frac{1}{2}$
- b) 14
- c) $3 \frac{1}{2}$
- d) 15

8. Terry has 4 hats and Dan has 5 more hats than Terry. How many hats does Dan have?

RA-PT-5

- a) 4
- b) 5
- c) 1
- d) 9

9. Solve: $\cdot 2 = 8$

SE-PT-2

- a) 6
- b) 2
- c) 8
- d) 4

10. Solve: $\frac{1}{3} \cdot \frac{3}{5} =$

RN-PT-7

a) $\frac{4}{8}$

b) $\frac{4}{15}$

c) $\frac{3}{15}$

d) $\frac{3}{8}$

11. $\frac{2}{4} + \frac{3}{2}$ has the same value as:

RN-PT-4

a) $2 + \frac{3}{4} + 2$

b) $\frac{5}{8} + \frac{5}{6}$

c) $\frac{5}{6}$

d) $\frac{3}{2} + \frac{2}{4}$

12. Solve: $\frac{2}{3} \cdot \frac{3}{2} =$

RN-PT-3

a) $\frac{5}{6}$

b) $\frac{6}{6}$

c) $2\frac{1}{6}$

d) $\frac{6}{5}$

13. Which of the following fractions is equal to 1?

PA-RN-1

a) $\frac{21}{21}$

b) $1\frac{2}{3}$

c) $\frac{12}{22}$

d) $\frac{11}{1}$

14. Simplify $6(3 - 2)$

PA-PT-6

a) 16

b) 6

c) 11

d) $6 + 3 - 2$

15. $3(12 + 2)$ has the same value as:

PA-PT-2

a) $3(12) \cdot (2)$

b) 36

c) $3 \cdot 12 \cdot 2$

d) $(12 + 2) + (12 + 2) + (12 + 2)$

16. $6x + 0 = 4x - 3$ is equivalent to:

PA-PT-7

a) $6x = 4x - 3$

b) $6x + 0 = 4x$

c) $6x + 4x = 10x$

d) $6x - 3 = 0$

17. Which of the following shows $\frac{6}{32}$ in its simplest form?

RN-PT-5

a) $\frac{6}{16}$

b) $\frac{3}{8}$

c) $\frac{3}{16}$

d) $\frac{6}{8}$

18. $20 + x$ is the same as:

SE-PT-6

a) $x + 20$

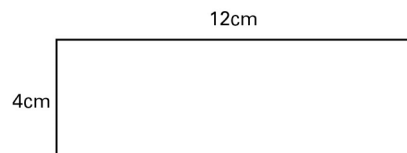
b) $20x$

c) $x - 20$

d) $\frac{x}{20}$

19. What is the perimeter of this rectangle?

RA-PT-1



a) 32 cm

b) 48 cm

c) 16 cm

d) 12 cm

20. A rectangular school playground measures 120 feet by 200 feet. What is the perimeter of the playground?

RA-PT-7

- a) 640 ft
- b) 320 ft
- c) 2400 ft
- d) 840 ft

21. $2xy$ has the same value as:

SE-PT-3

- a) $xy + xy + xy$
- b) $xy \div 2$
- c) $xy + xy$
- d) $(2 + x) + y$

22. $a \cdot (2 \cdot 3)$ has the same value as:

PA-PT-3

- a) $2a + 3$
- b) $a + 2 + 3$
- c) $(a \cdot 2) \cdot 3$
- d) $a \cdot 2 \cdot a \cdot 3$

23. $(4 \div 6) \div 2$ has the same value as:

PA-PT-4

- a) $4 \cdot 6 \cdot 2$
- b) $4 \div 6 \cdot \frac{1}{2}$
- c) $4 \div (6 \div 2)$
- d) $4 \div 6 \cdot 2$

24. $\frac{3}{0}$ has the same value as:

RN-PT-2

- a) 0
- b) undefined
- c) $3 \cdot 0$
- d) 3

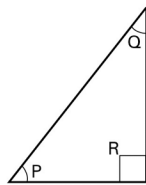
25. If m is 3, what is the value of $5 \cdot m - 3$?

SE-PT-5

- a) 2
- b) 12
- c) 15
- d) 0

26. The triangle shown here is a right triangle, and angles P and Q are equal. What is angle Q?

RA-PT-6



- a) 45°
- b) 90°
- c) 25°
- d) 75°

27. $\frac{4x}{4}$ has the same value as:

SE-PT-4

- a) 1
- b) x
- c) 4
- d) $16x$

28. If $p = 3$, what is $6p$?

SE-PT-7

- a) 63
- b) 9
- c) 6
- d) 18

Appendix C

Student ID # _____

Powersource

Answer each question below. Circle your answer.

1. What do you need to add to eighty-three to make one hundred?
2. Write the fraction $\frac{3}{9}$ in its simplest form.
3. There were two thousand people at a concert. Nine hundred and ninety-two of them were women. How many of the people were not women?
4. Write a fraction that is less than $\frac{4}{9}$.

5. Write a fraction that has a denominator of 100 **and** is equivalent to $\frac{7}{20}$.

6. What value of x makes the equation true?

$$x - 9 = 32$$

- a) 23
- b) 41
- c) 32
- d) 9

7. Solve: $6n = 36$

- a) 12
- b) 2
- c) 30
- d) 6

8. What is the next step to solve this equation?

$$x - 7 = 13$$

- a) Subtract 7 from both sides
- b) Add x to both sides
- c) Add 7 to both sides
- d) Subtract 13 from both sides

9. Write a different fraction that is equivalent to three-fifths.

10. $b = 14 + a$. When a equals 7, what is the value of b ?

11. If $\frac{12}{n} = \frac{36}{21}$, then n equals:

- a) 3
- b) 7
- c) 36
- d) 63

12. Which of the following ratios is equivalent to the ratio of 6 to 4?

- a) 12 to 18
- b) 12 to 8
- c) 8 to 6
- d) 4 to 6
- e) 2 to 3

13. Charlie can type 32 words per minute. At this rate, how long would it take him, in minutes, to type 128 words?

- a) 1
- b) 3
- c) 4
- d) 2

14. Sam's uncle is 21 years older than Sam. His uncle is 42. What equation could you use to solve for Sam's age, s ?

- a) $s + 21 = 42$
- b) $\frac{42}{21} = s$
- c) $s - 21 = 42$
- d) $s - 42 = 21$

15. Which of the following shows the distributive property being used correctly to simplify the expression: $3(4) + 3(2)$

- a) $3(4)(2)$
- b) $3(4 + 2)$
- c) $4(3 + 2)$
- d) $4(3) + 2(3)$

16. What is the value of p in the equation below ?

$$\frac{1}{4}p = 4$$

- a) $p = 4$
- b) $p = 16$
- c) $p = 4 \frac{1}{4}$
- d) $p = 3 \frac{3}{4}$

17. For all numbers k ,

$k + k + k + k + k$ can be written as

- a) $k + 5$
 - b) $5k$
 - c) k^5
 - d) $5(k + 1)$
18. Which of the following is equal to $6(x + 6)$?

- a) $x + 12$
- b) $6x + 6$
- c) $6x + 12$
- d) $6x + 36$
- e) $6x + 66$

19. Simplify using the distributive property.

$$y(y - 6) =$$

20. How much change will John get back from \$5.00 if he buys 2 notebooks that cost \$1.80 each?

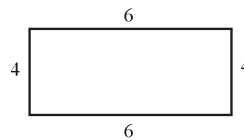
- a) \$1.40
- b) \$2.40
- c) \$3.20
- d) \$3.60

21. The perimeter of a square is 36 inches. What is the length of one side of the square?

- a) 4 inches
- b) 6 inches
- c) 9 inches
- d) 18 inches

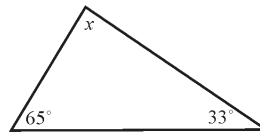
22. Which of the following numerical expressions gives the area of the rectangle below?

- a) $4 \cdot 6$
- b) $4 + 6$
- c) $2(4 \cdot 6)$
- d) $2(4 + 6)$
- e) $4 + 6 + 4 + 6$



23. What is the value of x in the triangle?

- a) 65°
- b) 82°
- c) 90°
- d) 92°
- e) 98°



24. If $3 + w = b$, then $w =$

- a) $\frac{3}{9}$
- b) $b \cdot 3$
- c) $b + 3$
- d) $3 - b$
- e) $b - 3$

25. In which list of fractions are all of the fractions equivalent?

- a) $\frac{1}{2}, \frac{2}{4}, \frac{4}{6}$
- b) $\frac{2}{3}, \frac{4}{6}, \frac{8}{12}$
- c) $\frac{2}{5}, \frac{4}{10}, \frac{8}{50}$
- d) $\frac{3}{4}, \frac{4}{6}, \frac{6}{8}$

26. n is a number. When n is multiplied by 7, and 6 is then added, the result is 41.
Which of these equations represents this relation?

- a) $7n + 6 = 41$
- b) $7n + - 6 = 41$
- c) $7n \cdot 6 = 41$
- d) $7(n + 6) = 41$

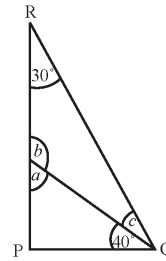
27. Explain why the fraction $\frac{\frac{1}{2}}{\frac{3}{4}}$ is equivalent to the fraction $\frac{2}{3}$?

28. What would be your answer if you were asked to multiply $8 \cdot (x + \frac{3}{4})$?

- a) $8x + \frac{3}{4}$
- b) $8 \frac{3}{4}x$
- c) $8x + 6$
- d) $x + 6$

29. The diagram shows triangle PQR.

What are the sizes of the angles a , b , and c ?



Appendix D

Appendix Table D1. Descriptive Statistics of Pretest scores by District and Treatment

Design	District	Treatment	<i>N</i>	Pretest mean	<i>SD</i>	Min	Max
Between	AZ-1	Treatment	93	17.40	4.49	8.00	27.00
	CA-3	Control	133	17.75	4.22	5.00	25.00
		Treatment	62	18.48	4.82	7.00	27.00
	CA-4	Control	93	17.15	5.18	6.00	26.00
		Treatment	186	17.24	5.18	6.00	26.00
	CA-6	Control	407	17.28	4.88	5.00	28.00
		Treatment	501	17.73	4.77	1.00	27.00
Within	CA-1	Control	349	17.37	4.38	2.00	27.00
		Treatment	523	19.40	3.78	8.00	27.00
	CA-2	Control	349	18.33	4.35	2.00	27.00
		Treatment	421	19.26	4.04	3.00	28.00
	CA-5	Control	422	17.24	4.28	4.00	26.00
		Treatment	552	18.45	4.71	1.00	27.00

Appendix Table D2. Descriptive Statistics of Posttest scores by District and Treatment

Design	District	Treatment	<i>N</i>	Posttest mean	<i>SD</i>	Min	Max
Between	AZ-1	Treatment	93	17.80	6.23	5.00	33.00
	CA-3	Control	133	18.89	4.75	2.00	29.00
		Treatment	62	19.39	6.67	6.00	33.00
	CA-4	Control	93	17.12	5.54	7.00	31.00
		Treatment	186	18.12	5.81	3.00	31.00
	CA-6	Control	407	16.09	5.57	3.00	32.00
		Treatment	501	16.56	6.47	1.00	31.00
Within	CA-1	Control	349	17.26	5.28	4.00	31.00
		Treatment	523	20.09	6.18	5.00	33.00
	CA-2	Control	349	16.36	5.44	1.00	31.00
		Treatment	421	17.75	5.69	2.00	31.00
	CA-5	Control	422	17.70	5.35	3.00	30.00
		Treatment	552	18.74	6.19	0.00	33.00

Appendix Table D3. Descriptive Statistics of Pretest Scores by School

Design	District	School	Treatment	<i>N</i>	Pretest mean	<i>SD</i>	Min	Max
Between	AZ-1	#1	Treatment	93	17.40	4.49	8.00	27.00
	CA-3	#2	Control	72	17.61	4.37	5.00	25.00
		#3	Treatment	62	18.48	4.82	7.00	27.00
		#4	Control	61	17.92	4.06	9.00	25.00
	CA-4	#5	Control	51	15.88	4.75	6.00	26.00
		#6	Treatment	45	18.29	4.20	9.00	25.00
		#7	Treatment	58	13.17	4.58	6.00	22.00
		#8	Treatment	83	19.52	4.36	6.00	26.00
		#9	Control	42	18.69	5.32	7.00	26.00
	CA-6	#10	Control	52	17.06	3.77	9.00	23.00
		#11	Control	169	16.47	4.99	6.00	27.00
		#12	Control	151	17.81	5.27	5.00	28.00
		#13	Treatment	322	17.13	4.81	1.00	26.00
		#14	Treatment	179	18.80	4.53	6.00	27.00
		#15	Control	35	19.26	3.01	10.00	26.00
Within	CA-1	#16	Control	62	17.65	4.02	8.00	26.00
			Treatment	226	19.25	3.67	8.00	27.00
		#17	Control	152	16.74	4.73	2.00	26.00
			Treatment	114	18.90	4.04	10.00	26.00
		#18	Control	135	17.96	4.05	6.00	27.00
			Treatment	183	19.89	3.72	10.00	27.00
	CA-2	#19	Control	134	19.46	4.35	6.00	27.00
			Treatment	251	19.92	3.75	9.00	28.00
		#20	Control	215	17.62	4.21	2.00	27.00
			Treatment	170	18.27	4.26	3.00	26.00
	CA-5	#21	Treatment	41	14.76	4.57	5.00	24.00
		#22	Control	134	16.87	4.42	4.00	26.00
			Treatment	142	17.16	4.98	1.00	26.00
		#23	Control	31	18.52	3.45	8.00	25.00
			Treatment	30	17.47	4.52	8.00	24.00
		#24	Control	132	16.02	4.23	5.00	25.00
			Treatment	187	20.92	3.62	10.00	27.00
		#25	Control	125	18.61	3.95	8.00	25.00
			Treatment	152	17.81	4.35	6.00	26.00

Appendix Table D4. Descriptive Statistics of Posttest Scores by School

Design	District	School	Treatment	<i>N</i>	Posttest mean	<i>SD</i>	Min	Max
Between	AZ-1	#1	Treatment	93	17.80	6.23	5.00	33.00
	CA-3	#2	Control	72	19.24	4.93	2.00	29.00
		#3	Treatment	62	19.39	6.67	6.00	33.00
		#4	Control	61	18.49	4.54	9.00	27.00
	CA-4	#5	Control	51	16.45	5.27	7.00	31.00
		#6	Treatment	45	18.02	5.71	3.00	28.00
		#7	Treatment	58	18.50	6.58	6.00	31.00
		#8	Treatment	83	17.90	5.33	3.00	29.00
		#9	Control	42	17.93	5.82	9.00	29.00
	CA-6	#10	Control	52	14.52	3.95	7.00	24.00
		#11	Control	169	14.94	5.71	3.00	30.00
		#12	Control	151	17.36	5.62	6.00	32.00
		#13	Treatment	322	15.36	6.16	1.00	29.00
		#14	Treatment	179	18.72	6.46	5.00	31.00
		#15	Control	35	18.54	4.72	8.00	26.00
Within	CA-1	#16	Control	62	17.03	5.58	7.00	31.00
			Treatment	226	19.55	5.87	5.00	33.00
		#17	Control	152	16.59	5.38	4.00	28.00
			Treatment	114	19.11	6.79	5.00	32.00
		#18	Control	135	18.11	4.95	5.00	28.00
			Treatment	183	21.36	5.96	5.00	33.00
	CA-2	#19	Control	134	16.55	5.55	1.00	30.00
			Treatment	251	17.29	5.88	2.00	31.00
		#20	Control	215	16.24	5.37	2.00	31.00
			Treatment	170	18.43	5.35	4.00	28.00
	CA-5	#21	Treatment	41	13.44	5.31	4.00	24.00
		#22	Control	134	16.66	5.31	5.00	27.00
			Treatment	142	18.56	6.35	4.00	31.00
		#23	Control	31	18.68	4.55	6.00	27.00
			Treatment	30	18.70	7.21	0.00	33.00
		#24	Control	132	16.77	4.89	3.00	28.00
			Treatment	187	22.09	4.61	9.00	32.00
		#25	Control	125	19.57	5.55	7.00	30.00
			Treatment	152	16.24	5.53	0.00	29.00

Appendix Table D5. Descriptive Statistics of Pretest Scores by Teacher in Between-School Design

Design	District	School	Treatment	Teacher ID	N	Pretest mean	SD	Min	Max
Between	AZ-1	#1	Treatment	124	27	16.33	3.85	10.00	24.00
				125	66	17.83	4.69	8.00	27.00
	CA-3	#2	Control	406	72	17.61	4.37	5.00	25.00
				401	25	20.00	4.31	12.00	27.00
		#3	Treatment	402	27	18.63	4.22	11.00	26.00
				404	10	14.30	5.54	7.00	23.00
				405	61	17.92	4.06	9.00	25.00
	CA-4	#5	Control	509	26	14.46	4.14	6.00	24.00
				511	25	17.36	4.97	8.00	26.00
		#6	Treatment	501	22	19.64	4.03	9.00	25.00
				502	23	17.00	4.02	10.00	24.00
		#7	Treatment	504	42	13.40	4.83	6.00	22.00
				515	16	12.56	3.92	7.00	21.00
				506	30	18.47	4.53	9.00	26.00
		#8	Treatment	507	26	20.12	4.18	6.00	25.00
				508	27	20.11	4.28	8.00	25.00
		#9	Control	512	11	21.55	3.80	15.00	25.00
				513	31	17.68	5.46	7.00	26.00
	CA-6	#10	Control	802	52	17.06	3.77	9.00	23.00
				812	60	20.07	4.08	10.00	27.00
		#11	Control	813	57	15.18	4.29	7.00	24.00
				814	52	13.75	4.23	6.00	23.00
		#12	Control	831	59	14.32	4.66	5.00	23.00
				832	92	20.04	4.37	8.00	28.00
		#13	Treatment	806	96	17.84	4.35	5.00	25.00
				807	62	15.77	4.46	1.00	24.00
				808	7	6.43	0.98	5.00	8.00
				809	30	17.57	4.26	5.00	25.00
				810	28	15.11	4.68	6.00	23.00
				834	99	18.48	4.54	3.00	26.00
				818	22	16.45	4.33	9.00	25.00
		#14	Treatment	819	53	16.87	4.21	6.00	25.00
				820	44	22.80	2.25	18.00	27.00
				821	52	17.46	3.75	7.00	23.00
				822	8	24.75	1.16	23.00	26.00
				816	35	19.26	3.01	10.00	26.00

Appendix Table D6. Descriptive Statistics of Pretest Scores by Teacher in Within-School (W-S) Design

Design	District	School	Treatment	Teacher ID	N	Pretest mean	SD	Min	Max
Within	CA-1	#16	Control	210	36	16.81	3.73	8.00	22.00
				229	26	18.81	4.20	9.00	26.00
			Treatment	207	80	17.98	3.33	9.00	24.00
				211	45	17.18	3.58	9.00	25.00
				218	45	19.47	3.33	8.00	26.00
				225	56	22.57	1.70	19.00	27.00
		#17	Control	208	52	18.31	4.84	5.00	26.00
				230	28	18.36	3.87	8.00	23.00
				231	44	14.20	5.19	2.00	24.00
				233	28	16.18	2.16	11.00	21.00
			Treatment	216	31	23.23	2.29	16.00	26.00
				232	83	17.29	3.30	10.00	24.00
		#18	Control	201	53	18.09	4.87	6.00	27.00
				205	23	18.43	2.25	14.00	22.00
				228	59	17.64	3.80	8.00	26.00
			Treatment	202	59	23.24	1.99	19.00	27.00
				203	34	18.56	3.53	13.00	25.00
				209	90	18.19	3.18	10.00	24.00
	CA-2	#19	Control	308	20	16.95	5.93	6.00	25.00
				309	97	19.97	3.75	7.00	27.00
				310	17	19.53	4.67	11.00	26.00
			Treatment	305	83	20.45	3.36	11.00	27.00
				306	97	20.27	3.75	12.00	28.00
				307	71	18.85	4.01	9.00	26.00
		#20	Control	303	99	17.31	4.21	2.00	25.00
				304	116	17.89	4.22	5.00	27.00
			Treatment	301	92	18.00	4.47	3.00	26.00
				302	41	17.29	3.89	10.00	25.00
				312	37	20.03	3.69	9.00	26.00

(table continues)

Appendix Table D6. Descriptive Statistics of Pretest Scores by Teacher in W-S Design (*continued*)

Design	District	School	Treatment	Teacher ID	N	Pretest mean	SD	Min	Max
Within	CA-5	#21	Treatment	601	41	14.76	4.57	5.00	24.00
		#22		604	43	15.23	4.76	4.00	24.00
			Control	606	39	15.49	3.73	8.00	23.00
				609	52	19.27	3.52	10.00	26.00
				602	38	18.82	4.14	9.00	25.00
		Treatment		605	38	16.58	4.50	8.00	25.00
				607	34	17.91	4.73	6.00	26.00
				608	32	15.09	5.95	1.00	24.00
		#23	Control	610	31	18.52	3.45	8.00	25.00
			Treatment	611	30	17.47	4.52	8.00	24.00
		#24		614	20	16.40	4.28	7.00	23.00
			Control	616	112	15.96	4.24	5.00	25.00
				615	101	22.45	2.41	15.00	27.00
			Treatment	617	86	19.13	3.98	10.00	26.00
		#25		618	7	12.57	3.64	8.00	17.00
			Control	620	72	19.69	3.38	11.00	25.00
				621	46	17.83	3.87	8.00	25.00
				619	59	16.12	4.58	6.00	26.00
			Treatment	622	68	17.78	3.71	6.00	24.00
				623	25	21.88	2.37	16.00	26.00

Appendix Table D7. Descriptive Statistics of Posttest Scores by Teacher in Between-School Design

Design	District	School	Treatment	Teacher ID	N	Posttest Mean	SD	Min	Max
Between	AZ-1	#1	Treatment	124	27	13.96	4.43	5.00	21.00
				125	66	19.36	6.21	5.00	33.00
	CA-3	#2	Control	406	72	19.24	4.93	2.00	29.00
				401	25	21.36	6.30	11.00	33.00
		#3	Treatment	402	27	18.11	6.92	6.00	29.00
				404	10	17.90	6.28	6.00	25.00
				405	61	18.49	4.54	9.00	27.00
	CA-4	#5	Control	509	26	14.35	4.26	7.00	22.00
				511	25	18.64	5.39	11.00	31.00
		#6	Treatment	501	22	17.27	6.04	6.00	28.00
				502	23	18.74	5.41	3.00	26.00
		#7	Treatment	504	42	19.33	6.81	9.00	31.00
				515	16	16.31	5.53	6.00	24.00
		#8	Treatment	506	30	15.73	5.26	3.00	26.00
				507	26	17.69	4.51	10.00	25.00
				508	27	20.52	5.17	10.00	29.00
		#9	Control	512	11	22.00	5.46	13.00	29.00
				513	31	16.48	5.30	9.00	27.00
	CA-6	#10	Control	802	52	14.52	3.95	7.00	24.00
				812	60	18.08	6.43	3.00	30.00
		#11	Control	813	57	13.35	4.24	6.00	22.00
				814	52	13.06	4.66	5.00	26.00
				831	59	15.86	4.96	6.00	28.00
		#12	Control	832	92	18.32	5.84	6.00	32.00
				806	96	16.14	6.17	2.00	29.00
				807	62	13.58	5.52	4.00	28.00
		#13	Treatment	808	7	8.29	3.64	4.00	14.00
				809	30	16.20	5.13	7.00	25.00
				810	28	13.43	6.64	1.00	24.00
				834	99	16.53	6.24	1.00	29.00
				818	22	16.00	4.81	6.00	25.00
		#14	Treatment	819	53	15.30	5.46	5.00	27.00
				820	44	26.27	3.05	18.00	31.00
				821	52	16.06	4.43	6.00	28.00
				822	8	24.63	2.67	21.00	29.00
		#15	Control	816	35	18.54	4.72	8.00	26.00

Appendix Table D8. Descriptive Statistics of Posttest Scores by Teacher in Within-School (W-S) Design

Design	District	School	Treatment	Teacher ID	N	Posttest mean	SD	Min	Max
Within	CA-1	#16	Control	210	36	14.08	4.25	7.00	23.00
				229	26	21.12	4.56	13.00	31.00
			Treatment	207	80	18.98	4.75	8.00	29.00
				211	45	15.09	5.30	5.00	26.00
				218	45	18.16	4.94	7.00	26.00
				225	56	25.09	4.15	15.00	33.00
		#17	Control	208	52	16.50	6.29	5.00	26.00
				230	28	18.79	4.91	11.00	28.00
				231	44	14.00	4.62	4.00	22.00
				233	28	18.64	2.91	12.00	24.00
			Treatment	216	31	27.19	2.06	23.00	32.00
				232	83	16.10	5.31	5.00	26.00
		#18	Control	201	53	16.64	5.76	5.00	28.00
				205	23	17.91	4.34	8.00	26.00
				228	59	19.51	3.98	11.00	28.00
			Treatment	202	59	27.07	2.63	21.00	33.00
				203	34	16.38	5.48	7.00	27.00
				209	90	19.50	4.73	5.00	29.00
	CA-2	#19	Control	308	20	11.85	6.82	1.00	25.00
				309	97	17.35	4.73	5.00	30.00
				310	17	17.53	5.83	8.00	26.00
			Treatment	305	83	17.10	5.15	7.00	30.00
				306	97	18.04	6.11	3.00	31.00
				307	71	16.51	6.29	2.00	28.00
		#20	Control	303	99	17.91	5.17	2.00	30.00
				304	116	14.82	5.15	5.00	31.00
			Treatment	301	92	20.39	4.38	8.00	28.00
				302	41	14.95	5.26	4.00	26.00
				312	37	17.41	5.54	8.00	28.00

(table continues)

Appendix Table D8. Descriptive Statistics of Posttest Scores by Teacher in W-S Design (*continued*)

Design	District	School	Treatment	Teacher ID	N	Posttest mean	SD	Min	Max
Within	CA-5	#21	Treatment	601	41	13.44	5.31	4.00	24.00
		#22		604	43	14.40	3.94	6.00	21.00
			Control	606	39	14.56	5.37	5.00	25.00
				609	52	20.12	4.40	11.00	27.00
				602	38	20.74	6.38	7.00	30.00
		Treatment		605	38	19.50	6.56	5.00	30.00
				607	34	18.74	6.55	4.00	31.00
				608	32	14.66	3.87	7.00	25.00
		#23	Control	610	31	18.68	4.55	6.00	27.00
			Treatment	611	30	18.70	7.21	0.00	33.00
		#24		614	20	16.10	4.56	7.00	23.00
			Control	616	112	16.88	4.96	3.00	28.00
				615	101	22.50	4.04	9.00	32.00
			Treatment	617	86	21.60	5.18	10.00	31.00
		#25		618	7	10.00	2.65	7.00	14.00
			Control	620	72	22.08	4.23	9.00	30.00
				621	46	17.09	5.00	8.00	25.00
				619	59	13.59	4.02	6.00	22.00
			Treatment	622	68	16.04	5.03	0.00	28.00
				623	25	23.00	4.27	12.00	29.00